b. Explain why the results indicate that there may be a problem of positive autocorrelation. Can you give arguments why, in economic models, positive autocorrelation is more likely than negative autocorrelation?

c. What are the effects of autocorrelation on the properties of the OLS estimator? Think about unbiasedness, consistency and the BLUE property.

d. Describe two different approaches to handle the autocorrelation problem in the above case. Which one would you prefer?

From now on, assume that S_t and Y_t are nonstationary $I(1)$ series.

e. Are there indications that the relationship between the two variables is 'spurious'?

f. Explain what we mean by 'spurious regressions'.

g. Are there indications that there is a cointegrating relationship between S_t and Y_t?

h. Explain what we mean by a 'cointegrating relationship'.

i. Describe two different tests that can be used to test the null hypothesis that S_t and Y_t are not cointegrated.

j. How do you interpret the coefficient estimate of 0.098 under the hypothesis that S_t and Y_t are cointegrated?

k. Are there reasons to correct for autocorrelation in the error term when we estimate a cointegrating regression?

l. Explain intuitively why the estimator for a cointegrating parameter is superconsistent.

m. Assuming that S_t and Y_t are cointegrated, describe what we mean by an error-correction mechanism. Give an example. What do we learn from it?

n. How can we consistently estimate an error-correction model?

Exercise 9.3 (Cointegration – Empirical)

In the files INCOME we find quarterly data on UK nominal consumption and income, for 1971:1 to 1985:2 ($T = 58$). Part of these data was used in Exercise 8.3.

a. Test for a unit root in the consumption series using several augmented Dickey–Fuller tests.

b. Perform a regression by OLS explaining consumption from income. Test for cointegration using two different tests.

c. Perform a regression by OLS explaining income from consumption. Test for cointegration.

d. Compare the estimation results and R^2s from the last two regressions.

e. Determine the error-correction term from one of the two regressions and estimate an error-correction model for the change in consumption. Test whether the adjustment coefficient is zero.

f. Repeat the last question for the change in income. What do you conclude?

10 Models Based on Panel Data

A panel data set contains repeated observations over the same units (individuals, households, firms), collected over a number of periods. Although panel data are typically collected at the micro-economic level, it has become increasingly common to pool individual time series of a number of countries or industries and analyse them simultaneously. The availability of repeated observations on the same units allows economists to specify and estimate more complicated and more realistic models than a single cross-section or a single time series would do. The disadvantages are more of a practical nature: because we repeatedly observe the same units, it is usually no longer appropriate to assume that different observations are independent. This may complicate the analysis, particularly in nonlinear and dynamic models. Furthermore, panel data sets very often suffer from missing observations. Even if these observations are missing in a random way (see below), the standard analysis has to be adjusted.

This chapter provides an introduction to the analysis of panel data. A simple linear panel data model is presented in Section 10.1, and some advantages compared with cross-sectional or time series data are discussed in the context of this model. Section 10.2 pays attention to the so-called fixed effects and random effects models, and discusses issues relating to the choice between these two basic models. An empirical illustration is provided in Section 10.3. The introduction of a lagged dependent variable in the linear model complicates consistent estimation, and, as will be discussed in Section 10.4, instrumental variables procedures or GMM provide interesting alternatives. Section 10.5 provides an empirical example on the estimation of a partial adjustment model for a firm's capital structure. Increasingly, panel data approaches are used in a macro-economic context to investigate the dynamic properties of economic variables. Section 10.6 discusses the recent literature on unit root and cointegration tests in heterogeneous panels. In micro-economic applications, the model of interest often involves limited dependent variables, and panel data extensions of logit, probit and tobit models are discussed in Section 10.7. The problems associated with incomplete panel data and selection bias are discussed in Section 10.8, while Section 10.9 concludes this chapter with a discussion on pseudo panel data and repeated cross-sections. Extensive discussions of the econometrics of panel data can be found in Wooldridge (2002), Hsiao (2003), Arellano (2003), Baltagi (2005) and Cameron and Trivedi (2005).

10.1 Introduction to Panel Data Modeling

An important advantage of panel data compared with time series or cross-sectional data sets is that they allow identification of certain parameters or questions, without the need to make restrictive assumptions. For example, panel data make it possible to analyse *changes* on an *individual* level. Consider a situation in which the average consumption level rises by 2% from one year to another. Panel data can identify whether this rise is the result of, for example, an increase of 2% for all individuals or an increase of 4% for approximately one-half of the individuals and no change for the other half (or any other combination). That is, panel data are not only suitable to model or explain why individual units behave differently but also to model why a given unit behaves differently at different time periods (for example, because of a different past).

We shall, below, index all variables with an i for the individual[1] ($i = 1, \ldots, N$) and a t for the time period ($t = 1, \ldots, T$). The standard linear regression model can then be written as

$$y_{it} = \beta_0 + x_{it}'\beta + \varepsilon_{it}, \tag{10.1}$$

where x_{it} is a K-dimensional vector of explanatory variables, which – for reasons that will become clear below – does not contain an intercept term.[2] This model imposes that the intercept β_0 and the slope coefficients in β are identical for all individuals and time periods. The error term in (10.1) varies over individuals and time and captures all unobservable factors that affect y_{it}. To estimate this model by OLS, the usual conditions are required to achieve unbiasedness, consistency or efficiency; see Chapters 2, 4 and 5. For example, if $E\{\varepsilon_{it}\} = 0$ and $E\{x_{it}\varepsilon_{it}\} = 0$, the OLS estimator is consistent for β_0 and β under weak regularity conditions. Given that we repeatedly observe the same individuals, however, it is typically unrealistic to assume that the error terms from different periods are uncorrelated. For example, a person's wage will be affected by unobservable characteristics that vary little over time. As a result, routinely computed standard errors for OLS, based on the assumption of i.i.d. error terms, tend to be misleading in panel data applications. Moreover, OLS is likely to be inefficient relative to an estimator that exploits the correlation over time in ε_{it}.

A very frequently employed panel data model assumes that

$$\varepsilon_{it} = \alpha_i + u_{it}, \tag{10.2}$$

where u_{it} is assumed to be homoskedastic and not correlated over time. The component α_i is time invariant and homoskedastic across individuals. The model specified by

[1] While we refer to the cross-sectional units as individuals, they could also refer to other units like firms, countries, industries, households or assets.

[2] The elements in β are indexed as β_1 to β_K, where the first element, unlike in the previous chapters, does not refer to the intercept.

(10.1)–(10.2) is referred to as an error components or **random effects model**, and we shall discuss it in more detail below. Estimation by (feasible) generalized least squares exploiting the imposed error structure (which implies that the serial correlation in ε_{it} can completely be attributed to α_i) typically leads to a more efficient estimator for β_0 and β than ordinary least squares.

The assumption that $E\{x_{it}\varepsilon_{it}\} = 0$ states that the observable regressors in x_{it} are uncorrelated with the unobservable characteristics in both α_i and u_{it}. This means that the explanatory variables are exogenous. In many applications this assumption is considered restrictive, and there are reasons to believe that $E\{x_{it}\alpha_i\} \neq 0$. That is, the unobserved heterogeneity in α_i is correlated with one or more of the explanatory variables. For example, in a wage equation a person's unobserved ability is likely to affect wages (y_{it}), but also a person's education level (included in x_{it}). In a firm-level investment equation, unobserved firm characteristics may affect investment decisions (y_{it}) as well characteristics in x_{it} (e.g. the cost of capital). In a cross-sectional context, the standard approach to handle this problem is the use of instrumental variables (see Chapter 5). With panel data, it is possible to exploit the particular nature of the data owing to the availability of repeated observations on the same individuals.

In a **fixed effects model**, this problem is addressed by including an individual-specific intercept term in the model. In this case, we write the model as

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it}, \tag{10.3}$$

where α_i ($i = 1, \ldots, N$) are fixed unknown constants that are estimated along with β, and where u_{it} is typically assumed to be i.i.d. over individuals and time. The overall intercept term β_0 is omitted, as it is subsumed by the individual intercepts α_i. It is common to refer to α_i as fixed (individual) effects. The fixed effects α_i capture all (un)observable time-invariant differences across individuals. In this approach, consistent estimation does not impose that α_i and x_{it} are uncorrelated.

The possibility of treating the α_is as fixed parameters has some great advantages, but also some disadvantages. Most panel data models are estimated under either the fixed effects or the random effects assumption, and we shall discuss this extensively in Section 10.2. First, the next two subsections discuss some potential advantages of panel data in more detail.

10.1.1 Efficiency of Parameter Estimators

Because panel data sets are typically larger than cross-sectional or time series data sets, and explanatory variables vary over two dimensions (individuals and time) rather than one, estimators based on panel data are quite often more accurate than from other sources. Even with identical sample sizes, the use of a panel data set will often yield more efficient estimators than a series of independent cross-sections (where different units are sampled in each period). To illustrate this, consider the following special case of the random effects model in (10.1)–(10.2) where we only include time dummies, i.e.

$$y_{it} = \mu_t + \alpha_i + u_{it}, \tag{10.4}$$

where each μ_t is an unknown parameter corresponding to the population mean in period t. Suppose we are not interested in the mean μ_t in a particular period, but in the change of μ_t from one period to another. In general the variance of the efficient estimator for $\mu_t - \mu_s (s \neq t)$, $\hat{\mu}_t - \hat{\mu}_s$, is given by

$$V\{\hat{\mu}_t - \hat{\mu}_s\} = V\{\hat{\mu}_t\} + V\{\hat{\mu}_s\} - 2\,\mathrm{cov}\{\hat{\mu}_t, \hat{\mu}_s\}$$

with $\hat{\mu}_t = 1/N \sum_{i=1}^{N} y_{it}$ $(t = 1, \ldots, T)$. Typically, if a panel data set is used, the covariance between $\hat{\mu}_t$ and $\hat{\mu}_s$ will be positive. For example, under the random effects assumptions of equation (10.2) it equals σ_α^2/N. However, if two independent cross-sectional data sets are used, different periods will contain different individuals, so $\hat{\mu}_t$ and $\hat{\mu}_s$ will have zero covariance. In other words, if one is interested in changes from one period to another, a panel will yield more efficient estimators than a series of cross-sections.

Note, however, that the reverse is also true, in the sense that repeated cross-sections will be more informative than a panel when one is interested in a sum or average of μ_t over several periods. At a more intuitive level, panel data may provide better information because the *same* individuals are repeatedly observed. On the other hand, having the same individuals rather than different ones may imply less variation in the explanatory variables and thus relatively inefficient estimators. A comprehensive analysis on the choice between a pure panel, a pure cross-section and a combination of these two data sources is provided in Nijman and Verbeek (1990). Their results indicate that, when exogenous variables are included in the model and one is interested in the parameters that measure the effects of these variables, a panel data set will typically yield more efficient estimators than a series of cross-sections with the same number of observations.

10.1.2 Identification of Parameters

A second advantage of the availability of panel data is that it reduces identification problems. Although this advantage may come under different headings, in many cases it involves identification in the presence of endogenous regressors or measurement error, robustness to omitted variables and the identification of individual dynamics.

Let us start with an illustration of the last of these. There are two alternative explanations for the often observed phenomenon that individuals who have experienced an event in the past are more likely to experience that event in the future. The first explanation is that the fact that an individual has experienced the event changes his or her preferences, constraints, etc., in such a way that he or her is more likely to experience that event in the future. The second explanation says that individuals may differ in unobserved characteristics that influence the probability of experiencing the event (but are not influenced by the experience of the event). Heckman (1978a) terms the former explanation 'true state dependence' and the latter 'spurious state dependence'. A well-known example concerns the 'event' of being unemployed. The availability of panel data will ease the problem of distinguishing between true and spurious state dependence, because individual histories are observed and can be included in the model.

Omitted variable bias arises if a variable that is correlated with the included variables is excluded from the model. A classical example is the estimation of production functions (Mundlak, 1961). In many cases, especially in the case of small firms, it is desirable to include management quality as an input in the production function.

In general, however, management quality is unobservable. Suppose that a production function of the Cobb–Douglas type is given by

$$y_{it} = \beta_0 + x_{it}'\beta + m_i\beta_{K+1} + u_{it}, \qquad (10.5)$$

where y_{it} denotes log output, x_{it} is a K-dimensional vector of log inputs, both for firm i at time t, and m_i denotes the management quality for firm i (which is assumed to be constant over time). The unobserved variable m_i is expected to be negatively correlated with the other inputs in x_{it}, since a high-quality management will probably result in a more efficient use of inputs. Therefore, unless $\beta_{K+1} = 0$, deletion of m_i from (10.5) will lead to biased estimates of the other parameters in the model. If panel data are available, this problem can be resolved by introducing a firm-specific effect $\alpha_i = \beta_0 + m_i\beta_{K+1}$ and considering this as a fixed unknown parameter. Note that without additional information it is not possible to identify β_{K+1}; a restriction that identifies β_{K+1} is the imposition of constant returns to scale.[3]

In a similar way, a fixed time effect can be included in the model to capture the effect of all (observed and unobserved) variables that do not vary over the individual units. This illustrates the proposition that panel data can reduce the effects of omitted variable bias, or – in other words – estimators from a panel data set may be more robust to an incomplete model specification.

Finally, in many cases panel data will provide 'internal' instruments for regressors that are endogenous or subject to measurement error. That is, transformations of the original variables can often be argued to be uncorrelated with the model's error term and correlated with the explanatory variables themselves and no external instruments are needed. For example, if x_{it} is correlated with α_i, it can be argued that $x_{it} - \bar{x}_i$, where $\bar{x}_i$ is the time average for individual i, is uncorrelated with α_i and provides a valid instrument for x_{it}. More generally, estimating the model under the fixed effects assumption eliminates α_i from the error term and, consequently, eliminates all endogeneity problems relating to it. This will be illustrated in the next section. An extensive discussion of the benefits and limitations of panel data is provided in Hsiao (1985).

10.2 The Static Linear Model

In this section we discuss the static linear model in a panel data setting. We start with the fixed effects model, and pay attention to the within estimator and the first-difference estimator. Next, we present the random effects model. Subsequently, we discuss the choice between fixed effects and random effects, as well as alternative estimation procedures that can be considered to be somewhere between a fixed effects and random effects treatment. This section also pays attention to goodness-of-fit, heteroskedasticity and autocorrelation, and to robust covariance matrix estimation.

10.2.1 The Fixed Effects Model

The fixed effects model is simply a linear regression model in which the intercept terms vary over the individual units i, i.e.

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it}, \quad u_{it} \sim IID(0, \sigma_u^2), \qquad (10.6)$$

[3] Constant returns to scale implies that $\beta_{K+1} = 1 - (\beta_1 + \cdots + \beta_K)$.

where it is usually assumed that all x_{it} are independent of all u_{it}. We can write this in the usual regression framework by including a dummy variable for each unit i in the model. That is,

$$y_{it} = \sum_{j=1}^{N} \alpha_j d_{ij} + x_{it}'\beta + u_{it}, \qquad (10.7)$$

where $d_{ij} = 1$ if $i = j$ and 0 elsewhere. We thus have a set of N dummy variables in the model. The parameters $\alpha_1, \ldots, \alpha_N$ and β can be estimated by ordinary least squares in (10.7). The implied estimator for β is referred to as the **least squares dummy variable (LSDV) estimator**. It may, however, be numerically unattractive to have a regression model with so many regressors. Fortunately one can compute the estimator for β in a simpler way. It can be shown that exactly the same estimator for β is obtained if the regression is performed in deviations from individual means. Essentially, this implies that we eliminate the individual effects α_i first by transforming the data. To see this, first note that

$$\bar{y}_i = \alpha_i + \bar{x}_i'\beta + \bar{u}_i, \qquad (10.8)$$

where $\bar{y}_i = T^{-1}\sum_t y_{it}$ and $\bar{x}_i$ and $\bar{u}_i$ are defined in a similar way. Consequently, we can write

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta + (u_{it} - \bar{u}_i). \qquad (10.9)$$

This is a regression model in deviations from individual means and does not include the individual effects α_i. The transformation that produces observations in deviations from individual means, as in (10.9), is called the **within transformation**. The OLS estimator for β obtained from this transformed model is often called the **within estimator** or **fixed effects estimator**, and it is exactly identical to the LSDV estimator described above. It is given by

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i). \qquad (10.10)$$

If it is assumed that *all* x_{it} are independent of *all* u_{it} (compare assumption (A2) from Chapter 2), the fixed effects estimator can be shown to be unbiased for β. If, in addition, normality of u_{it} is imposed, $\hat{\beta}_{FE}$ also has a normal distribution. For consistency,[4] it is required that

$$E\{(x_{it} - \bar{x}_i)u_{it}\} = 0 \qquad (10.11)$$

(compare assumption (A7) in Chapters 2 and 5). Sufficient for this is that x_{it} is uncorrelated with u_{it} and that $\bar{x}_i$ has no correlation with the error term. These conditions are in turn implied by

$$E\{x_{it}u_{is}\} = 0 \quad \text{for all } s, t, \qquad (10.12)$$

[4] Unless stated otherwise, we consider in this chapter consistency for the number of individuals N going to infinity. This corresponds to the common situation that we have panels with large N and small T.

in which case we call x_{it} **strictly exogenous**. A strictly exogenous variable is not allowed to depend upon current, future and past values of the error term. In some applications this may be restrictive. Clearly, it excludes the inclusion of lagged dependent variables in x_{it}, but any x_{it} variable that depends upon the history of y_{it} would also violate the condition. For example, if we are explaining labour supply of an individual, we may want to include years of experience in the model, while quite clearly experience depends upon the person's labour history.

With explanatory variables independent of all errors, the N intercepts are estimated unbiasedly as

$$\hat{\alpha}_i = \bar{y}_i - \bar{x}_i'\hat{\beta}_{FE}, \quad i = 1, \ldots, N.$$

Under assumption (10.11) these estimators are consistent for the fixed effects α_i provided T goes to infinity. The reason why $\hat{\alpha}_i$ is inconsistent for fixed T is clear: when T is fixed, the individual averages $\bar{y}_i$ and $\bar{x}_i$ do not converge to anything if the number of individuals increases.

The covariance matrix for the fixed effects estimator $\hat{\beta}_{FE}$, assuming that u_{it} is i.i.d. across individuals and time with variance σ_u^2, is given by

$$V\{\hat{\beta}_{FE}\} = \sigma_u^2\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'\right)^{-1}. \qquad (10.13)$$

Unless T is large, using the standard OLS estimate for the covariance matrix based upon the within regression in (10.9) will underestimate the true variance. The reason is that in this transformed regression the error covariance matrix is singular (as the T transformed errors of each individual add up to zero) and the variance of $u_{it} - \bar{u}_i$ is $(T-1)/T\sigma_u^2$ rather than σ_u^2. A consistent estimator for σ_u^2 is obtained from the sum of squared residuals from the within estimator, divided by $N(T-1)$. Defining

$$\hat{u}_{it} = y_{it} - \hat{\alpha}_i - x_{it}'\hat{\beta}_{FE} = y_{it} - \bar{y}_i - (x_{it} - \bar{x}_i')\hat{\beta}_{FE},$$

we estimate σ_u^2 as

$$\hat{\sigma}_u^2 = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{u}_{it}^2. \qquad (10.14)$$

It is possible to apply the usual degrees of freedom correction, in which case K is subtracted from the denominator. Note that using the standard OLS covariance matrix in model (10.7) with N individual dummies is reliable, because the degrees of freedom correction involves N additional unknown parameters corresponding to the individual intercept terms. Under weak regularity conditions, the fixed effects estimator is asymptotically normal, so that the usual inference procedures can be used (like t and Wald tests).

Essentially, the fixed effects model concentrates on differences 'within' individuals. That is, it is explaining to what extent y_{it} differs from $\bar{y}_i$ and does not explain why $\bar{y}_i$ is different from $\bar{y}_j$. The parametric assumptions about β, on the other hand, impose that a change in x has the same (ceteris paribus) effect, whether it is a change from one period to the other or a change from one individual to another. When interpreting the results, however, from a fixed effects regression, it may be important to realize that the parameters are identified only through the within dimension of the data.

10.2.2 *The First-difference Estimator*

An alternative way to eliminate the individual effects α_i is to first-difference equation (10.6). This results in

$$y_{it} - y_{i,t-1} = (x_{it} - x_{i,t-1})'\beta + (u_{it} - u_{i,t-1})$$

or

$$\Delta y_{it} = \Delta x_{it}'\beta + \Delta u_{it}, \tag{10.15}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$. Applying OLS to this equation yields the **first-difference estimator**

$$\hat{\beta}_{FD} = \left(\sum_{i=1}^{N}\sum_{t=2}^{T}\Delta x_{it}\Delta x_{it}'\right)^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}\Delta x_{it}\Delta y_{it}. \tag{10.16}$$

Consistency of this estimator requires that

$$E\{\Delta x_{it}\Delta u_{it}\} = 0$$

or

$$E\{(x_{it} - x_{i,t-1})(u_{it} - u_{i,t-1})\} = 0. \tag{10.17}$$

This condition is weaker than the strict exogeneity condition in (10.12). For example, it would allow correlation between x_{it} and $u_{i,t-2}$. To compute the standard errors for $\hat{\beta}_{FD}$, it should be taken into account that Δu_{it} exhibits serial correlation. While the conditions for consistency of the first-differences estimator are slightly weaker than those for the within estimator, it is, in general, somewhat less efficient. For $T = 2$, both estimators are identical (see Exercise 10.1). If the two estimators provide very different results, this suggests that assumption (10.12) is problematic.

A simple and sometimes attractive estimator is the **differences-in-differences estimator**. Because it is an intuitively attractive approach, it also helps us to understand the merits of panel data. Suppose we are interested in estimating the impact of a certain 'treatment' upon a given outcome variable (see Section 7.7). While the terminology comes from medical sciences, treatment may also refer to social or economic interventions, e.g. enrolment into a labour training programme, receipt of a transfer payment from a social programme or being a member of a trade union. A typical outcome variable is 'earnings'. Let the binary regressor of interest be

$$r_{it} = 1 \text{ if individual } i \text{ receives a treatment in period } t;$$
$$= 0 \text{ otherwise.}$$

Let us assume a fixed effects model for y_{it} as

$$y_{it} = \delta r_{it} + \mu_t + \alpha_i + u_{it},$$

where μ_t is a time-specific fixed effect. For simplicity, the only regressor is r_{it} (in addition to the time and individual fixed effects). In general, the impact of a treatment can be inferred from a comparison of people receiving treatment with those who do not and by a comparison of people before and after the treatment. Panel data combines both.

The individual effects can be eliminated by a first-difference transformation. That is,

$$\Delta y_{it} = \delta\Delta r_{it} + \Delta\mu_t + \Delta u_{it}. \tag{10.18}$$

Assuming that $E\{\Delta r_{it}\Delta u_{it}\} = 0$, the treatment effect δ can be estimated consistently by OLS of Δy_{it} upon Δr_{it} and a set of time dummies. Because the individual effects α_i are eliminated, this procedure allows correlation between α_i and the treatment indicator. This is important, because in many applications one can argue that individuals with certain (unobserved) characteristics are more likely to receive treatment (or to participate in some programme). Obviously, this approach is very similar to the fixed effects estimator, with the only difference that the first-difference transformation is employed rather than the within transformation.

Let us consider a situation in which there are only two time periods and individuals may receive a treatment in period 2. Thus $r_{i1} = 0$ for all i, while $r_{i2} = 1$ for a subset of the individuals. OLS in (10.18) corresponds to a regression of $y_{i2} - y_{i1}$ upon the treatment dummy and a constant (corresponding to the time effect). The resulting estimate for δ corresponds to the sample average of $y_{i2} - y_{i1}$ for the treated minus the average for the nontreated. Define $\Delta\bar{y}_{i2}^{treated}$ as the average for the treated ($r_{i2} = 1$) and $\Delta\bar{y}_{i2}^{nontreated}$ as the average for the nontreated ($r_{i2} = 0$). Then the OLS estimate is simply

$$\hat{\delta} = \Delta\bar{y}_{i2}^{treated} - \Delta\bar{y}_{i2}^{nontreated}.$$

This estimator is called the **differences-in-differences estimator**, because one estimates the time difference for the treated and untreated groups and then takes the difference between the two. The first-differencing takes care of unobservable fixed effects and controls for unobservable (time-invariant) differences between individuals (e.g. health status, ability, intelligence, ...). The second difference compares treated with untreated individuals. The formulation of the model in (10.18) makes clear that we need to assume that the time effects μ_t are common across treated and untreated individuals.

In economics the above methodology is often applied when the data arise from a natural experiment. A natural experiment occurs when some exogenous event (often a change in government policy or the passage of a law) changes the environment in which individuals, families or firms operate. A natural experiment always has a control group, which is not affected by the policy change, and a treatment group, which is thought to be affected by the policy change. Unlike with a true experiment where treatment and control groups can be randomly chosen, in a natural experiment these two groups arise from a particular policy change. In order to control for systematic differences between the control and treatment group, we need two periods of data, one before and one after the treatment. Thus the sample consists of four (sub)groups: the control group before and after the treatment and the treatment group before and after the treatment. Averages within these four subsamples are the building blocks of the differences-in-differences estimator; see Cameron and Trivedi (2005, Chapter 22) for more discussion.

10.2.3 The Random Effects Model

It is commonly assumed in regression analysis that all factors that affect the dependent variable, but that have not been included as regressors, can be appropriately summarized by a random error term. In our case, this leads to the assumption that the α_i are random factors, independently and identically distributed over individuals. Thus we write the random effects model as

$$y_{it} = \beta_0 + x_{it}'\beta + \alpha_i + u_{it}, \quad u_{it} \sim IID(0, \sigma_u^2); \quad \alpha_i \sim IID(0, \sigma_\alpha^2), \qquad (10.19)$$

where $\alpha_i + u_{it}$ is treated as an error term consisting of two components: an individual specific component, which does not vary over time, and a remainder component, which is assumed to be uncorrelated over time.[5] That is, all correlation of the error terms over time is attributed to the individual effects α_i. It is assumed that α_i and u_{it} are mutually independent and independent of x_{js} (for all j and s). This implies that the OLS estimator for β_0 and β from (10.19) is unbiased and consistent. The error components structure implies that the composite error term $\alpha_i + u_{it}$ exhibits a particular form of autocorrelation (unless $\sigma_\alpha^2 = 0$). Consequently, routinely computed standard errors for the OLS estimator are incorrect and a more efficient (GLS) estimator can be obtained by exploiting the structure of the error covariance matrix.

To derive the GLS estimator,[6] first note that for individual i all error terms can be stacked as $\alpha_i\iota_T + u_i$, where $\iota_T = (1, 1, \ldots, 1)'$ of dimension T and $u_i = (u_{i1}, \ldots, u_{iT})'$. The covariance matrix of this vector is (see Hsiao, 2003, Section 3.3)

$$V\{\alpha_i\iota_T + u_i\} = \Omega = \sigma_\alpha^2\iota_T\iota_T' + \sigma_u^2 I_T,$$

where I_T is the T-dimensional identity matrix. This can be used to derive the generalized least squares (GLS) estimator for the parameters in (10.19). For each individual, we can transform the data by premultiplying the vectors $y_i = (y_{i1}, \ldots, y_{iT})'$, etc., by Ω^{-1}, which is given by

$$\Omega^{-1} = \sigma_u^{-2}\left[I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2}\iota_T\iota_T'\right],$$

which can also be written as

$$\Omega^{-1} = \sigma_u^{-2}\left[\left(I_T - \frac{1}{T}\iota_T\iota_T'\right) + \psi\frac{1}{T}\iota_T\iota_T'\right],$$

[5] This model is sometimes referred to as a (one-way) error components model.
[6] It may be instructive to re-read the general introduction to GLS estimation in Section 4.2.

where

$$\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}.$$

Noting that $I_T - (1/T)\iota_T\iota_T'$ transforms the data in deviations from individual means and $(1/T)\iota_T\iota_T'$ takes individual means, the GLS estimator for β can be written as

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' + \psi T\sum_{i=1}^{N}(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'\right)^{-1}$$
$$\times\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) + \psi T\sum_{i=1}^{N}(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})\right), \qquad (10.20)$$

where $\bar{x} = (1/(NT))\sum_{i,t} x_{it}$ denotes the overall average of x_{it}. It is easy to see that for $\psi = 0$ the fixed effects estimator arises. Because $\psi \to 0$ if $T \to \infty$, it follows that the fixed and random effects estimators are equivalent for large T. If $\psi = 1$, the GLS estimator is just the OLS estimator (and Ω is diagonal). From the general formula for the GLS estimator it can be derived that

$$\hat{\beta}_{GLS} = \Delta\hat{\beta}_B + (I_K - \Delta)\hat{\beta}_{FE},$$

where

$$\hat{\beta}_B = \left(\sum_{i=1}^{N}(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'\right)^{-1}\sum_{i=1}^{N}(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})$$

is the so-called **between estimator** for β. It is the OLS estimator in the model for individual means

$$\bar{y}_i = \beta_0 + \bar{x}_i'\beta + \alpha_i + \bar{u}_i, \quad i = 1, \ldots, N. \qquad (10.21)$$

The matrix Δ is a weighting matrix and is proportional to the inverse of the covariance matrix of $\hat{\beta}_B$ (see Hsiao, 2003, Section 3.4, for details). That is, the GLS estimator is a matrix-weighted average of the between estimator and the within estimator, where the weight depends upon the relative variances of the two estimators. (The more accurate one gets, the higher the weight.)

The between estimator effectively discards the time series information in the data set. The GLS estimator, under the current assumptions, is the optimal combination of the within estimator and the between estimator, and is therefore more efficient than either of these two estimators. The OLS estimator (with $\psi = 1$) is also a linear combination of the two estimators, but not the efficient one. Thus, GLS will be more efficient than OLS, as usual. If the explanatory variables are independent of all u_{it} and all α_i, the GLS estimator is unbiased. It is a consistent estimator for N or T or both, tending to infinity if, in addition to (10.11), it also holds that $E\{\bar{x}_i u_{it}\} = 0$ and most importantly that

$$E\{\bar{x}_i\alpha_i\} = 0. \qquad (10.22)$$

Note that these conditions are also required for the between estimator to be consistent.

An easy way to compute the GLS estimator is obtained by noting that it can be determined as the OLS estimator in a transformed model (compare Chapter 4), given by

$$(y_{it} - \vartheta \bar{y}_i) = \beta_0(1 - \vartheta) + (x_{it} - \vartheta \bar{x}_i)'\beta + v_{it}, \quad (10.23)$$

where $\vartheta = 1 - \psi^{1/2}$. The error term in this transformed regression is i.i.d. over individuals and time. Note again that $\psi = 0$ corresponds to the within estimator ($\vartheta = 1$). In general, a fixed proportion ϑ of the individual means is subtracted from the data to obtain this transformed model ($0 \leq \vartheta \leq 1$).

Of course, the variance components σ_α^2 and σ_u^2 are unknown in practice. In this case we can use the feasible GLS estimator (EGLS), where the unknown variances are consistently estimated in a first step. The estimator for σ_u^2 is easily obtained from the within residuals, as given in (10.14). For the between regression the error variance is $\sigma_\alpha^2 + (1/T)\sigma_u^2$, which we can estimate consistently by

$$\hat{\sigma}_B^2 = \frac{1}{N}\sum_{i=1}^{N}(\bar{y}_i - \hat{\beta}_{0B} - \bar{x}_i'\hat{\beta}_B)^2, \quad (10.24)$$

where $\hat{\beta}_{0B}$ is the between estimator for β_0. From this, a consistent estimator for σ_α^2 follows as

$$\hat{\sigma}_\alpha^2 = \hat{\sigma}_B^2 - \frac{1}{T}\hat{\sigma}_u^2. \quad (10.25)$$

Again, it is possible to adjust this estimator by applying a degrees of freedom correction, implying that the number of regressors $K + 1$ is subtracted in the denominator of (10.24) (see Hsiao, 2003, Section 3.3). The resulting EGLS estimator is referred to as the **random effects estimator** for β (and β_0), denoted below as $\hat{\beta}_{RE}$. It is also known as the Balestra–Nerlove estimator.

Under weak regularity conditions, the random effects estimator is asymptotically normal. Its covariance matrix is given by

$$V\{\hat{\beta}_{RE}\} = \sigma_u^2\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' + \psi T\sum_{i=1}^{N}(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'\right)^{-1}, \quad (10.26)$$

which shows that the random effects estimator is more efficient than the fixed effects estimator as long as $\psi > 0$. The gain in efficiency is due to the use of the between variation in the data ($\bar{x}_i - \bar{x}$). The covariance matrix in (10.26) is routinely estimated by the OLS expressions in the transformed model (10.23).

In summary, we have seen a range of estimators for the parameter vector β. The basic two are:

1. The **between estimator**, exploiting the between dimension of the data (differences between individuals), determined as the OLS estimator in a regression of individual averages of y on individual averages of x (and a constant). Consistency, for $N \to \infty$, requires that $E\{\bar{x}_i\alpha_i\} = 0$ and $E\{\bar{x}_i\bar{u}_i\} = 0$. Typically this means that the explanatory variables are strictly exogenous and uncorrelated with the individual specific effect α_i.

2. The **fixed effects (within) estimator**, exploiting the within dimension of the data (differences within individuals), determined as the OLS estimator in a regression in deviations from individual means. It is consistent for β for $T \to \infty$ or $N \to \infty$, provided that $E\{(x_{it} - \bar{x}_i)u_{it}\} = 0$. Again this requires the x variables to be strictly exogenous, but it does not impose any restrictions upon the relationship between α_i and x_{it}.

Two estimators that combine the within and between dimension of the data are:

3. The **OLS estimator**, exploiting both dimensions (within and between) but not efficiently. Determined (of course) as OLS in the original model given in (10.19). Consistency for $T \to \infty$ or $N \to \infty$ requires that $E\{x_{it}(u_{it} + \alpha_i)\} = 0$. This requires the explanatory variables to be uncorrelated with α_i but does not impose that they are strictly exogenous. It suffices that x_{it} and u_{it} are contemporaneously uncorrelated.
4. The **random effects (EGLS) estimator**, combining the information from the between and within dimensions in an efficient way. It is consistent for $T \to \infty$ or $N \to \infty$ under the combined conditions of 1 and 2. It can be determined as a weighted average of the between and within estimator or as the OLS estimator in a regression where the variables are transformed as $y_{it} - \hat{\vartheta}\bar{y}_i$, where $\hat{\vartheta}$ is an estimate for $\vartheta = 1 - \psi^{1/2}$ with $\psi = \sigma_u^2/(\sigma_u^2 + T\sigma_\alpha^2)$.

Further, we have also considered:

5. The **first-difference (FD) estimator**, determined as OLS after first-differencing the equation of interest. This estimator is an alternative to the fixed effects estimator based on the within transformation, and it also only exploits the time variation in the data. Consistency requires that $E\{(x_{it} - x_{i,t-1})(u_{it} - u_{i,t-1})\} = 0$. If u_{it} is i.i.d., the first-difference estimator is less efficient than the within estimator; for $T = 2$ they are identical.

10.2.4 Fixed Effects or Random Effects?

The choice between a fixed effects and a random effects approach is not easy, and in many applications, particularly when T is small, the differences in the estimates for β appear to be substantial. The most common view is that the discussion should not be about the 'true nature' of the effects α_i. The appropriate interpretation is that the fixed effects approach is conditional upon the values for α_i. That is, it essentially considers the distribution of y_{it} given α_i, where the α_is can be estimated. This makes sense intuitively if the individuals in the sample are 'one of a kind', and cannot be viewed as a random draw from some underlying population. This interpretation is probably most appropriate when i denotes countries, (large) companies or industries, and predictions we want to make are for a particular country, company or industry. Inferences are thus with respect to the effects that are in the sample.

In contrast, the random effects approach is not conditional upon the individual α_is, but 'integrates them out'. In this case, we are usually not interested in the particular value of some person's α_i; we just focus on arbitrary individuals who have certain characteristics. The random effects approach allows one to make inference with respect to the population characteristics. One way to formalize this is noting that the random effects model states that

$$E\{y_{it}|x_{it}\} = x_{it}'\beta,$$

while the fixed effects model estimates

$$E\{y_{it}|x_{it}, \alpha_i\} = x_{it}'\beta + \alpha_i.$$

Note that the β coefficients in these two conditional expectations are the same only if $E\{\alpha_i|x_{it}\} = 0$. Accordingly, a first reason why one may prefer the fixed effects estimator is that some interest lies in α_i, which makes sense if the number of units is relatively small and of a specific nature. That is, identification of individual units is important.

However, even if we are interested in the larger population of individual units, and a random effects framework seems appropriate, the fixed effects estimator may be preferred. The reason for this is that it may be the case that α_i and x_{it} are correlated, in which case the random effects approach, ignoring this correlation, leads to inconsistent estimators. We saw an example of this above, where α_i included management quality and was argued to be correlated with the other inputs included in the production function. The problem of correlation between the individual effects α_i and the explanatory variables in x_{it} can be handled by using the fixed effects approach, which essentially eliminates the α_i from the model, and thus eliminates any problems that they may cause.

Hausman (1978) has suggested a test for the null hypothesis that x_{it} and α_i are uncorrelated. The general idea of a **Hausman test** is that two estimators are compared: one that is consistent under both the null and alternative hypothesis and one that is consistent (and typically efficient) under the null hypothesis only. A significant difference between the two estimators indicates that the null hypothesis is unlikely to hold. In the present case, assume that $E\{u_{it}x_{is}\} = 0$ for all s, t, so that the fixed effects estimator $\hat{\beta}_{FE}$ is consistent for β irrespective of the question as to whether x_{it} and α_i are uncorrelated, while the random effects estimator $\hat{\beta}_{RE}$ is consistent and efficient only if x_{it} and α_i are not correlated. Let us consider the difference vector $\hat{\beta}_{FE} - \hat{\beta}_{RE}$. To evaluate the significance of this difference, we need its covariance matrix. In general this would require us to estimate the covariance between $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$, but, because the latter estimator is efficient under the null hypothesis, it can be shown that (under the null)

$$V\{\hat{\beta}_{FE} - \hat{\beta}_{RE}\} = V\{\hat{\beta}_{FE}\} - V\{\hat{\beta}_{RE}\}. \tag{10.27}$$

Consequently, we can compute the Hausman test statistic as

$$\xi_H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[\hat{V}\{\hat{\beta}_{FE}\} - \hat{V}\{\hat{\beta}_{RE}\}]^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE}), \tag{10.28}$$

where the $\hat{V}$s denote estimates of the true covariance matrices. Under the null hypothesis, which implicitly says that $\text{plim}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) = 0$, the statistic ξ_H has an asymptotic Chi-squared distribution with K degrees of freedom, where K is the number of elements in β.

The Hausman test thus tests whether the fixed effects and random effects estimators are significantly different. Computationally, this is relatively easy because the covariance matrix satisfies (10.27). An important reason why the two estimators would be different is the existence of correlation between x_{it} and α_i, although other sorts of misspecification can also lead to rejection (we shall see an example of this below). A practical problem when computing (10.28) is that the covariance matrix in square brackets may not be positive definite in finite samples, such that its inverse cannot be computed. As an alternative, it is possible to test for a subset of the elements in β.

10.2.5 Goodness-of-fit

The computation of goodness-of-fit measures in panel data applications is somewhat uncommon. One reason is the fact that one may attach different importance to explaining the within and between variation in the data. Another reason is that the usual R^2 or adjusted R^2 criteria are only appropriate if the model is estimated by OLS.

Our starting point here is the definition of the R^2 in terms of the squared correlation coefficient between actual and fitted values, as presented in Section 2.4. This definition has the advantage that it produces values within the [0, 1] interval, irrespective of the estimator that is used to generate the fitted values. Recall that it corresponds to the standard definition of the R^2 (in terms of sums of squares) if the model is estimated by OLS (provided that an intercept term is included). In the current context, the total variation in y_{it} can be written as the sum of the within variation and the between variation, that is

$$\frac{1}{NT}\sum_{i,t}(y_{it} - \bar{y})^2 = \frac{1}{NT}\sum_{i,t}(y_{it} - \bar{y}_i)^2 + \frac{1}{N}\sum_{i}(\bar{y}_i - \bar{y})^2,$$

where $\bar{y}$ denotes the overall sample average. Now, we can define alternative versions of an R^2 measure, depending upon the dimension of the data that we are interested in.

For example, the fixed effects estimator is chosen to explain the within variation as well as possible, and thus maximizes the 'within R^2' given by

$$R^2_{within}(\hat{\beta}_{FE}) = \text{corr}^2\{\hat{y}_{it}^{FE} - \hat{y}_i^{FE}, y_{it} - \bar{y}_i\}, \tag{10.29}$$

where $\hat{y}_{it}^{FE} - \hat{y}_i^{FE} = (x_{it} - \bar{x}_i)'\hat{\beta}_{FE}$ and corr^2 denotes the squared correlation coefficient. The between estimator, being an OLS estimator in the model in terms of individual means, maximizes the 'between R^2', which we define as

$$R^2_{between}(\hat{\beta}_B) = \text{corr}^2\{\hat{y}_i^B, \bar{y}_i\}, \tag{10.30}$$

where $\hat{y}_i^B = \bar{x}_i'\hat{\beta}_B$. The OLS estimator maximizes the overall goodness-of-fit and thus the overall R^2, which is defined as

$$R^2_{overall}(\hat{\beta}) = \text{corr}^2\{\hat{y}_{it}, y_{it}\}, \tag{10.31}$$

with $\hat{y}_{it} = x_{it}'b$. It is possible to define within, between and overall R^2s for an arbitrary estimator $\hat{\beta}$ for β by using as fitted values $\hat{y}_{it} = x_{it}'\hat{\beta}$, $\hat{y}_i = (1/T)\sum_t \hat{y}_{it}$ and $\hat{y} = (1/(NT))\sum_{i,t}\hat{y}_{it}$, where the intercept terms are omitted (and irrelevant).[7] For the

[7] These definitions correspond to the R^2 measures as computed in Stata.

fixed effects estimator this ignores the variation captured by the $\hat{\alpha}_i$s. If we take into account the variation explained by the N estimated intercepts $\hat{\alpha}_i$, the fixed effects model perfectly fits the between variation. This is somewhat unsatisfactory, though, as it is hard to argue that the fixed effects $\hat{\alpha}_i$ *explain* the variation between individuals, they just capture it. Put differently, if we ask ourselves: why does individual i consume on average more than another individual, the answer provided by $\hat{\alpha}_i$ is simply: because it is individual i. Given this argument, and because the $\hat{\alpha}_i$s are often not computed, it seems appropriate to ignore this part of the model.

Taking the definition in terms of the squared correlation coefficients, the three measures above can be computed for any of the estimators that we considered. If we take the random effects estimator, which is (asymptotically) the most efficient estimator if the assumptions of the random effects model are valid, the within, between and overall R^2s are necessarily smaller than for the fixed effects, between and OLS estimator respectively. This, again, stresses that goodness-of-fit measures are not adequate to choose between alternative estimators. They provide, however, possible criteria for choosing between alternative (potentially non-nested) specifications of the model.

10.2.6 Alternative Instrumental Variables Estimators

The fixed effects estimator eliminates anything that is time invariant from the model. This may be a high price to pay for allowing the x variables to be correlated with the individual specific heterogeneity α_i. For example, we may be interested in the effect of time-invariant variables (like gender) on a person's wage. Actually, there is no need to restrict attention to the fixed and random effects assumptions only, as it is possible to derive instrumental variables estimators that can be considered to be in between a fixed and random effects approach.

To see this, let us first of all note that we can write the fixed effects estimator as

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)(y_{it}-\bar{y}_i)$$
$$= \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)x_{it}'\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)y_{it}. \qquad (10.32)$$

Writing the estimator like this shows that it has the interpretation of an instrumental variables estimator[8] for β in the model

$$y_{it} = \beta_0 + x_{it}'\beta + \alpha_i + u_{it},$$

where each explanatory variable is instrumented by its value in deviation from the individual specific mean. That is, x_{it} is instrumented by $x_{it} - \bar{x}_i$. Note that $E\{(x_{it} - \bar{x}_i)\alpha_i\} = 0$ by construction (if we take expectations over i and t), so that the IV estimator is consistent provided $E\{(x_{it} - \bar{x}_i)u_{it}\} = 0$, which is implied by the strict exogeneity of x_{it}. Clearly, if a particular element in x_{it} is known to be uncorrelated with α_i there is no need to instrument it; that is, this variable can be used as its own instrument. This route may also allow us to estimate the effect of time-invariant variables.

[8] It may be instructive to re-read Section 5.3 for a general discussion of instrumental variables estimation.

To describe the general approach, let us consider a linear model with four groups of explanatory variables (Hausman and Taylor, 1981):

$$y_{it} = \beta_0 + x_{1,it}'\beta_1 + x_{2,it}'\beta_2 + w_{1i}'\gamma_1 + w_{2i}'\gamma_2 + \alpha_i + u_{it}, \qquad (10.33)$$

where the x variables are time varying and the w variables are time invariant. The variables with index 1 are assumed to be uncorrelated with both α_i and all u_{is}. The variables $x_{2,it}$ and w_{2i} are correlated with α_i but not with any u_{is}. Under these assumptions, the fixed effects estimator would be consistent for β_1 and β_2, but would not identify the coefficients for the time-invariant variables. Moreover, it is inefficient because $x_{1,it}$ is needlessly instrumented. Hausman and Taylor (1981) suggest that (10.33) be estimated by instrumental variables using the following variables as instruments: $x_{1,it}$, w_{1i} and $x_{2,it} - \bar{x}_{2i}$, $\bar{x}_{1i}$. That is, the exogenous variables serve as their own instruments, $x_{2,it}$ is instrumented by its deviation from individual means (as in the fixed effects approach) and w_{2i} is instrumented by the individual average of $x_{1,it}$. Obviously, identification requires that the number of variables in $x_{1,it}$ is at least as large as that in w_{2i}. The resulting estimator, the **Hausman–Taylor estimator**, allows us to estimate the effect of time-invariant variables, even though the time-varying regressors are correlated with α_i. The trick here is to use the time averages of those time-varying regressors that are uncorrelated with α_i as instruments for the time-invariant regressors. Clearly, this requires that sufficient time-varying variables are included that have no correlation with α_i. Of course, it is a straightforward extension to include additional instruments in the procedure that are not based on variables included in the model. This is what one is forced to do in the cross-sectional case, where no transformations are available that can be argued to produce valid instruments. The strong advantage of the Hausman–Taylor approach is that one does not have to use external instruments. With sufficient assumptions, instruments can be derived within the model. Despite this important advantage, the Hausman–Taylor estimator plays a minor role in empirical work. A notable exception is Chowdhury and Nickell (1985).

Hausman and Taylor also show that the instrument set is equivalent to using $x_{1,it} - \bar{x}_{1i}$, $x_{2,it} - \bar{x}_{2i}$ and $x_{1,it}$, w_{1i}. This follows directly from the fact that taking different linear combinations of the original instruments does not affect the estimator. Hausman and Taylor also show how the nondiagonal covariance matrix of the error term in (10.33) can be exploited to improve the efficiency of the estimator. Nowadays, this would typically be handled in a GMM framework, as we shall see in the next section (see Arellano and Bover, 1995).

Two subsequent papers try to improve upon the efficiency of the Hausman–Taylor instrumental variables estimator by proposing a larger set of instruments. Amemiya and MaCurdy (1986) suggest the use of the *time-invariant* instruments $x_{1,i1} - \bar{x}_{1i}$ up to $x_{1,iT} - \bar{x}_{1i}$. This requires that $E\{(x_{1,it} - \bar{x}_{1i})\alpha_i\} = 0$ for *each t*. This assumption makes sense if the correlation between α_i and $x_{1,it}$ is due to a time-invariant component in $x_{1,it}$, such that $E\{x_{1,it}\alpha_i\}$ for a given t does not depend upon t. Breusch, Mizon and Schmidt (1989) nicely summarize this literature and suggest as additional instruments the use of the time-invariant variables $x_{2,i1} - \bar{x}_{2i}$ up to $x_{2,iT} - \bar{x}_{2i}$.

10.2.6 Robust Inference

Both the random effects and the fixed effects models assume that the presence of α_i captures all correlation between the unobservables in different time periods. That is, u_{it} is assumed to be uncorrelated over individuals and time. Provided that the x_{it} variables are strictly exogenous, the presence of autocorrelation in u_{it} does not result in inconsistency of the standard estimators. It does, however, invalidate the standard errors and resulting tests, just as we saw in Chapter 4. Moreover, it implies that the estimators are no longer efficient. For example, if the true covariance matrix Ω does have an error components structure, the random effects estimator no longer corresponds to the feasible GLS estimator for β. As we know, the presence of heteroskedasticity in u_{it} or – for the random effects model – in α_i has similar consequences.

One way to avoid misleading inferences, without the need to impose alternative assumptions on the structure of the covariance matrix Ω, is the use of the OLS, random effects or fixed effects estimators for β, while adjusting their standard errors for general forms of heteroskedasticity and autocorrelation. Consider the model[9]

$$y_{it} = x_{it}'\beta + \varepsilon_{it}, \tag{10.34}$$

without the assumption that ε_{it} has an error components structure. Consistency of the (pooled) OLS estimator

$$b = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}x_{it}'\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}y_{it} \tag{10.35}$$

for β requires that

$$E\{x_{it}\varepsilon_{it}\} = 0. \tag{10.36}$$

Assuming that error terms of different individuals are uncorrelated ($E\{\varepsilon_{it}\varepsilon_{js}\} = 0$ for all $i \neq j$), the OLS covariance matrix can be estimated by a variant of the Newey–West estimator from Chapter 4, given by

$$\hat{V}\{b\} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}x_{it}'\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} e_{it}e_{is}x_{it}x_{is}' \left(\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}x_{it}'\right)^{-1}, \tag{10.37}$$

where e_{it} denotes the OLS residual. This estimator allows for general forms of heteroskedasticity as well as arbitrary autocorrelation (within a given individual). Accordingly, (10.37) is referred to as a **panel-robust** estimate for the covariance matrix of the pooled OLS estimator. It is also known as a **cluster-robust** covariance matrix (where the identifier i indexing the individuals is the cluster variable). In a similar fashion, it is also possible to construct a robust estimator for the covariance matrix of the random effects estimator $\hat{\beta}_{RE}$ using the transformed model in (10.23). Note that the random effects estimator is not the appropriate EGLS estimator under these more general conditions.

When the model is estimated by the fixed effects estimator, a robust covariance matrix is obtained in a similar way, by replacing the regressors x_{it} in (10.37) with their within transformed counterparts, $\tilde{x}_{it} = x_{it} - \bar{x}_i$, and the OLS residuals with the residuals from the within regression (Arellano, 1987). That is,

$$\hat{V}\{\hat{\beta}_{FE}\} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T} \hat{u}_{it}\hat{u}_{is}\tilde{x}_{it}\tilde{x}_{is}' \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{x}_{it}\tilde{x}_{it}'\right)^{-1}, \tag{10.38}$$

where $\hat{u}_{it} = y_{it} - \hat{\alpha}_i - x_{it}'\hat{\beta}_{FE}$ denotes the within residual. For the first-difference estimator $\hat{\beta}_{FD}$ the first-differenced variables are employed (and the summation is from $t, s = 2$ to T). The above covariance matrix estimators are consistent for fixed T and $N \to \infty$ under weak regularity conditions. Bertrand, Duflo and Mullainathan (2004) provide a critical discussion on the computation of standard errors for the differences-in-differences estimator and, among other things, conclude that the panel-robust approach works reasonably well for moderate N. Similarly, Petersen (2006) advocates the use of panel-robust standard errors clustered by firms for sufficiently large N. If, on the other hand, N is small and $T \to \infty$, consistency can be achieved by using Bartlett weights in (10.37) as discussed in Subsection 4.10.2; see Arellano (2003, Section 2.3) for more details.

If one is willing to make specific assumptions about the form of heteroskedasticity or autocorrelation, it is possible to improve upon the efficiency of the OLS, random effects or fixed effects estimators by exploiting the structure of the error covariance matrix using a feasible GLS or maximum likelihood approach. An overview of a number of such estimators, which are typically computationally unattractive, is provided in Baltagi (2005, Chapter 5). Kmenta (1986) suggests a relatively simple feasible GLS estimator that allows for first-order autocorrelation in ε_{it} combined with individual specific heteroskedasticity, but does not allow for a time-invariant component in ε_{it} (see Baltagi, 2005, Section 10.4). Kiefer (1980) proposes a GLS estimator for the fixed effects model that allows for arbitrary covariances between u_{it} and u_{is}; see Arellano (2003, Section 2.3) or Hsiao (2003, Section 3.8) for more details. Wooldridge (2002, Subsection 10.4.3) describes a feasible GLS estimator where the covariance matrix Ω is estimated unrestrictedly from the pooled OLS residuals. Consistency of this estimator basically requires the same conditions as required by the random effects estimator, but it does not impose the error components structure. When N is sufficiently large relative to T, this feasible GLS estimator may provide an attractive alternative to the random effects approach.

10.2.7 Testing for Heteroskedasticity and Autocorrelation

Most of the tests that can be used for heteroskedasticity or autocorrelation in the random effects model are computationally burdensome. For the fixed effects model, which is essentially estimated by OLS, things are relatively less complex. Fortunately, as the fixed effects estimator can be applied even if we make the random effects assumption that α_i is i.i.d. and independent of the explanatory variables, the tests for the fixed effects model can also be used in the random effects case.

A fairly simple test for autocorrelation in the fixed effects model is based upon the Durbin–Watson test discussed in Chapter 4. The alternative hypothesis is that

$$u_{it} = \rho u_{i,t-1} + v_{it}, \tag{10.39}$$

[9] For notational convenience, the constant is assumed to be included in x_{it}.

where v_{it} is i.i.d. across individuals and time. This allows for autocorrelation over time with the restriction that each individual has the same autocorrelation coefficient ρ. The null hypothesis under test is H_0: $\rho = 0$ against the one-sided alternative $\rho < 0$ or $\rho > 0$. Let $\hat{u}_{it}$ denote the *residuals from the within regression* (10.9) or – equivalently – from (10.7). Then Bhargava, Franzini and Narendranathan (1983) suggest the following generalization of the Durbin–Watson statistic:

$$dw_p = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T}(\hat{u}_{it} - \hat{u}_{i,t-1})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{u}_{it}^2}. \quad (10.40)$$

Using similar derivations as Durbin and Watson, the authors are able to derive lower and upper bounds on the true critical values that depend upon N, T and K only. Unlike the true time series case, the inconclusive region for the panel data Durbin–Watson test is very small, particularly when the number of individuals in the panel is large. In Table 10.1 we present some selected lower and upper bounds for the true 5% critical values that can be used to test against the alternative of positive autocorrelation. The numbers in the table confirm that the inconclusive regions are small and also indicate that the variation with K, N or T is limited. In a model with three explanatory variables estimated over six time periods, we reject H_0: $\rho = 0$ at the 5% level if dw_p is smaller than 1.859 for $N = 100$ and 1.957 for $N = 1000$, both against the one-sided alternative of $\rho > 0$. For panels with very large N, Bhargava, Franzini and Narendranathan (1983) suggest simply to test if the computed statistic dw_p is less than two, when testing against positive autocorrelation. Because the fixed effects estimator is also consistent in the random effects model, it is also possible to use this panel data Durbin–Watson test in the latter model.

To test for heteroskedasticity in u_{it}, we can again use the fixed effects residuals $\hat{u}_{it}$. The auxiliary regression of the test regresses the squared within residuals $\hat{u}_{it}^2$ upon a constant and the J variables z_{it} that we think may affect heteroskedasticity. This is a variant of the Breusch–Pagan test[10] for heteroskedasticity discussed in Chapter 4. Its alternative hypothesis is that

$$V\{u_{it}\} = \sigma^2 h(z_{it}'\alpha), \quad (10.41)$$

Table 10.1 5% lower and upper bounds panel Durbin–Watson test

		$N = 100$		$N = 500$		$N = 1000$	
		d_L	d_U	d_L	d_U	d_L	d_U
$T = 6$	$K = 3$	1.859	1.880	1.939	1.943	1.957	1.959
	$K = 9$	1.839	1.902	1.935	1.947	1.954	1.961
$T = 10$	$K = 3$	1.891	1.904	1.952	1.954	1.967	1.968
	$K = 9$	1.878	1.916	1.949	1.957	1.965	1.970

Source: Bhargava, A., Franzini, L. and Narendranathan, W., (1983), Serial Correlation and the Fixed Effects Model, *The Review of Economic Studies* (**49**): 533-549. Reprinted by permission of Blackwell Publishing.

[10] In a panel data context, the term Breusch–Pagan test is usually associated with a Lagrange multiplier test in the random effects model for the null hypothesis that there are no individual specific effects ($\sigma_\alpha^2 = 0$); see Wooldridge (2002, Subsection 10.4.4) or Baltagi (2005, Subsection 4.2.1). In applications, this test almost always rejects the null hypothesis.

where h is an unknown continuously differentiable function with $h(0) = 1$, so that the null hypothesis that is tested is given by H_0: $\alpha = 0$. Under the null hypothesis, the test statistic, computed as $N(T-1)$ times the R^2 of the auxiliary regression, will have an asymptotic Chi-squared distribution, with J degrees of freedom. An alternative test can be computed from the residuals of the between regression and is based upon N times the R^2 of an auxiliary regression of the between residuals upon $\bar{z}_i$ or, more generally, upon $z_{i1}, \ldots, z_{iT}$. Under the null hypothesis of homoskedastic errors, the test statistic has an asymptotic Chi-squared distribution, with degrees of freedom equal to the number of variables included in the auxiliary regression (excluding the intercept). The alternative hypothesis of the latter test is less well defined.

10.3 Illustration: Explaining Individual Wages

In this section we shall apply a number of the above estimators when estimating an individual wage equation. The data[11] are taken from the Youth Sample of the National Longitudinal Survey held in the USA and comprise a sample of 545 full-time working males who completed their schooling by 1980 and were then followed over the period 1980–1987. The males in the sample were young, with an age in 1980 ranging from 17 to 23, and had entered the labour market fairly recently, with an average of 3 years of experience in the beginning of the sample period. The data and specifications we choose are similar to those in Vella and Verbeek (1998). Log wages are explained from years of schooling, years of experience and its square, dummy variables for being a union member, working in the public sector and being married and two racial dummies.

The estimation results[12] for the between estimator, based upon individual averages, and the within estimator, based upon deviations from individual means, are given in the first two columns of Table 10.2. First of all, it should be noted that the fixed effects or within estimator eliminates any time-invariant variables from the model. In this case, it means that the effects of schooling and race are wiped out. The differences between the two sets of estimates seem substantial, and we shall come back to this below. In the next column the OLS results are presented applied to the random effects model, where the standard errors are adjusted for heteroskedasticity and arbitrary forms of serial correlation based on the cluster-robust covariance matrix in (10.37). The last column presents the random effects EGLS estimator. As discussed in Subsection 10.2.3, the variances of the error components α_i and u_{it} can be estimated from the within and between residuals. In particular, we have $\hat{\sigma}_B^2 = 0.1209$ and $\hat{\sigma}_u^2 = 0.1234$. From this, we can consistently estimate σ_α^2 as $\hat{\sigma}_\alpha^2 = 0.1209 - 0.1234/8 = 0.1055$. Consequently, the factor ψ is estimated as

$$\hat{\psi} = \frac{0.1234}{0.1234 + 8 \times 0.1055} = 0.1276,$$

leading to $\hat{\vartheta} = 1 - \hat{\psi}^{1/2} = 0.6428$. This means that the EGLS estimator can be obtained from a transformed regression where 0.64 times the individual mean is subtracted

[11] The data used in this section are available as MALES.
[12] The estimation results in this section are obtained by Stata 9.2.

Table 10.2 Estimation results wage equation, males 1980–1987 (standard errors in parentheses)

Dependent variable: log(*wage*)

Variable	Between	Fixed effects	OLS	Random effects
constant	0.490	–	−0.034	−0.104
	(0.221)		(0.120)	(0.111)
schooling	0.095	–	0.099	0.101
	(0.011)		(0.009)	(0.009)
experience	−0.050	0.116	0.089	0.112
	(0.050)	(0.008)	(0.012)	(0.008)
experience^2	0.0051	−0.0043	−0.0028	−0.0041
	(0.0032)	(0.0006)	(0.0009)	(0.0006)
union member	0.274	0.081	0.180	0.106
	(0.047)	(0.019)	(0.028)	(0.018)
married	0.145	0.045	0.108	0.063
	(0.041)	(0.018)	(0.026)	(0.017)
black	−0.139	–	−0.144	−0.144
	(0.049)		(0.050)	(0.048)
hispanic	0.005	–	0.016	0.020
	(0.043)		(0.029)	(0.043)
public sector	−0.056	0.035	0.004	0.030
	(0.109)	(0.039)	(0.050)	(0.036)
within R^2	0.0470	0.1782	0.1679	0.1776
between R^2	0.2196	0.0006	0.2027	0.1835
overall R^2	0.1371	0.0642	0.1866	0.1808

from the original data. Recall that OLS imposes $\vartheta = 0$ while the fixed effects estimator employs $\vartheta = 1$. Note that both the OLS and the random effects estimates are in between the between and fixed effects estimates.

If the assumptions of the random effects model are satisfied, all four estimators in Table 10.2 are consistent, the random effects estimator being the most efficient one. If, however, the individual effects α_i are correlated with one or more of the explanatory variables, the fixed effects estimator is the only one that is consistent. This hypothesis can be tested by comparing the between and within estimators, or the within and random effects estimators, which leads to tests that are equivalent. The simplest one to perform is the Hausman test discussed in Subsection 10.2.4, based upon the latter comparison. The test statistic takes a value of 31.75 and reflects the differences in the coefficients on experience, experience squared and the union, married and public sector dummies. Under the null hypothesis, the statistic follows a Chi-squared distribution with five degrees of freedom, so that we have to reject the null at any reasonable level of significance.

Marital status is a variable that is likely to be correlated with the unobserved heterogeneity in α_i. Typically one would not expect an important *causal* effect of being married upon one's wage, so that the marital dummy is typically capturing other (unobservable) differences between married and unmarried workers. This is confirmed by the results in the table. If we eliminate the individual effects from the model and consider the fixed effects estimator, the effect of being married reduces to 4.5%, while for the between estimator, for example, it is almost 15%. Note that the effect of being

married in the fixed effects approach is identified only through people who change marital status over the sample period. Similar remarks can be made for the effect of union status upon a person's wage. Recall, however, that all estimators assume that the explanatory variables are uncorrelated with the idiosyncratic error term u_{it}. If such correlations were to exist, even the fixed effects estimator would be inconsistent. Vella and Verbeek (1998) concentrate on the impact of endogenous union status on wages for this group of workers and consider alternative, more complicated, estimators.

The goodness-of-fit measures confirm that the fixed effects estimator results in the largest within R^2 and thus explains the within variation as well as possible. The OLS estimator maximizes the usual (overall) R^2, while the random effects estimator results in reasonable R^2s in all dimensions. Recall that the OLS standard errors in Table 10.2 are adjusted for heteroskedasticity and arbitrary forms of serial correlation in the error terms. Routinely computed standard errors assuming i.i.d. error terms are inappropriate, and – in this application – sometimes less than half of the correct ones.

10.4 Dynamic Linear Models

Among the major advantages of panel data is the ability to model individual dynamics. Many economic models suggest that current behaviour depends upon past behaviour (persistence, habit formation, partial adjustment, etc.), so in many cases we would like to estimate a dynamic model on an individual level. The ability to do so is unique for panel data.

10.4.1 An Autoregressive Panel Data Model

Consider the linear dynamic model with exogenous variables and a lagged dependent variable, that is

$$y_{it} = x_{it}'\beta + \gamma y_{i,t-1} + \alpha_i + u_{it},$$

where it is assumed that u_{it} is $IID(0, \sigma_u^2)$. In the static model, we have seen arguments of consistency (robustness) and efficiency for choosing between a fixed or random effects treatment of the α_i. In a dynamic model the situation is substantially different, because $y_{i,t-1}$ will depend upon α_i, irrespective of the way we treat α_i. To illustrate the problems that this causes, we first consider the case where there are no exogenous variables included and the model reads

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + u_{it}, \quad |\gamma| < 1. \qquad (10.42)$$

Assume that we have observations on y_{it} for periods $t = 0, 1, \ldots, T$. Because $y_{i,t-1}$ and α_i are positively correlated, applying OLS to (10.42) is inconsistent, overestimating the true autoregressive coefficient (in the typical case where $\gamma > 0$). Similarly, the random effects approach is inconsistent.

The fixed effects estimator for γ is given by

$$\hat{\gamma}_{FE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \bar{y}_i)(y_{i,t-1} - \bar{y}_{i,-1})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1} - \bar{y}_{i,-1})^2}, \qquad (10.43)$$

where $\bar{y}_i = (1/T)\sum_{t=1}^{T} y_{it}$ and $\bar{y}_{i,-1} = (1/T)\sum_{t=1}^{T} y_{i,t-1}$. To analyse the properties of $\hat{\gamma}_{FE}$, we can substitute (10.42) into (10.43) to obtain

$$\hat{\gamma}_{FE} = \gamma + \frac{(1/(NT))\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it} - \bar{u}_i)(y_{i,t-1} - \bar{y}_{i,-1})}{(1/(NT))\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1} - \bar{y}_{i,-1})^2}. \quad (10.44)$$

This estimator, however, is biased and inconsistent for $N \to \infty$ and fixed T, as the last term on the right-hand side of (10.44) does not have expectation zero and does not converge to zero if N goes to infinity. In particular, it can be shown that (see Nickell, 1981; or Hsiao, 2003, Section 4.2)

$$\operatorname*{plim}_{N\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it} - \bar{u}_i)(y_{i,t-1} - \bar{y}_{i,-1}) = -\frac{\sigma_u^2}{T^2}\cdot\frac{(T-1) - T\gamma + \gamma^T}{(1-\gamma)^2} \neq 0. \quad (10.45)$$

Thus, for fixed T we have an inconsistent estimator. Note that this inconsistency is not caused by anything we assumed about the α_is, as these are eliminated in estimation. The problem is that the within transformed lagged dependent variable is correlated with the within transformed error. If $T \to \infty$, (10.45) converges to 0 so that the fixed effects estimator is consistent for γ if both $T \to \infty$ and $N \to \infty$.

One could think that the asymptotic bias for fixed T is quite small and therefore not a real problem. This is certainly not the case, as for finite T the bias can hardly be ignored. For example, if the true value of γ equals 0.5, it can easily be computed that (for $N \to \infty$)

$$\operatorname{plim}\hat{\gamma}_{FE} = -0.25 \quad \text{if } T = 2$$
$$\operatorname{plim}\hat{\gamma}_{FE} = -0.04 \quad \text{if } T = 3$$
$$\operatorname{plim}\hat{\gamma}_{FE} = 0.33 \quad \text{if } T = 10,$$

so even for moderate values of T the bias is substantial. Fortunately, there are relatively easy ways to avoid these biases.

To solve the inconsistency problem, we first of all start with a different transformation to eliminate the individual effects α_i, in particular we take first differences. This gives

$$y_{it} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + (u_{it} - u_{i,t-1}), \quad t = 2, \ldots, T. \quad (10.46)$$

If we estimate this by OLS, we do not obtain a consistent estimator for γ because $y_{i,t-1}$ and $u_{i,t-1}$ are, by definition, correlated, even if $T \to \infty$. In many applications, this first-difference estimator appears to be severely biased. However, this transformed specification suggests an instrumental variables approach. For example, $y_{i,t-2}$ is correlated with $y_{i,t-1} - y_{i,t-2}$ but not with $u_{i,t-1}$, unless u_{it} exhibits autocorrelation (which we excluded by assumption). This suggests an instrumental variables estimator[13] for γ as

$$\hat{\gamma}_{IV} = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T} y_{i,t-2}(y_{it} - y_{i,t-1})}{\sum_{i=1}^{N}\sum_{t=2}^{T} y_{i,t-2}(y_{i,t-1} - y_{i,t-2})}. \quad (10.47)$$

[13] See Section 5.3 for a general introduction to instrumental variables estimation.

A necessary condition for consistency of this estimator is that

$$\operatorname{plim}\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T}(u_{it} - u_{i,t-1})y_{i,t-2} = 0 \quad (10.48)$$

for T or N or both going to infinity. The estimator in (10.47) is one of the estimators proposed by Anderson and Hsiao (1981). They also proposed an alternative, where $y_{i,t-2} - y_{i,t-3}$ is used as an instrument. This gives

$$\hat{\gamma}_{IV}^{(2)} = \frac{\sum_{i=1}^{N}\sum_{t=3}^{T}(y_{i,t-2} - y_{i,t-3})(y_{it} - y_{i,t-1})}{\sum_{i=1}^{N}\sum_{t=3}^{T}(y_{i,t-2} - y_{i,t-3})(y_{i,t-1} - y_{i,t-2})}, \quad (10.49)$$

which is consistent (under regularity conditions) if

$$\operatorname{plim}\frac{1}{N(T-2)}\sum_{i=1}^{N}\sum_{t=3}^{T}(u_{it} - u_{i,t-1})(y_{i,t-2} - y_{i,t-3}) = 0. \quad (10.50)$$

Note that the second instrumental variables estimator requires an additional lag to construct the instrument, such that the effective number of observations used in estimation is reduced (one sample period is 'lost').

Consistency of both Anderson–Hsiao estimators is guaranteed by the assumption that u_{it} has no autocorrelation. However, Arellano (1989) has shown that the estimator that uses the first-differenced instrument, when exogenous variables are added to the model, suffers from large variances over a wide range of values for γ. In addition, Monte Carlo evidence by Arellano and Bover (1995) shows that the levels version of the Anderson–Hsiao estimator can have large biases and large standard errors, particularly when γ is close to one. Alternative estimators have been developed that build upon the Anderson–Hsiao approach. These approaches, formulated in a method of moments framework, unify the above estimators and eliminate the disadvantages of reduced sample sizes. The first step is to note that

$$\operatorname{plim}\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T}(u_{it} - u_{i,t-1})y_{i,t-2} = E\{(u_{it} - u_{i,t-1})y_{i,t-2}\} = 0 \quad (10.51)$$

is a moment condition (compare Chapter 5). Similarly,

$$\operatorname{plim}\frac{1}{N(T-2)}\sum_{i=1}^{N}\sum_{t=3}^{T}(u_{it} - u_{i,t-1})(y_{i,t-2} - y_{i,t-3})$$
$$= E\{(u_{it} - u_{i,t-1})(y_{i,t-2} - y_{i,t-3})\} = 0 \quad (10.52)$$

is a moment condition. Both IV estimators thus impose one moment condition in estimation. It is well known that imposing more moment conditions increases the efficiency of the estimators (provided the additional conditions are valid, of course). Arellano and Bond (1991) suggest that the list of instruments can be extended by exploiting additional moment conditions and letting their number vary with t. To do this, they keep T fixed. For example, when $T = 4$, we have

$$E\{(u_{i2} - u_{i1})y_{i0}\} = 0$$

as the moment condition for $t = 2$. For $t = 3$, we have

$$E\{(u_{i3} - u_{i2})y_{i1}\} = 0,$$

but it also holds that

$$E\{(u_{i3} - u_{i2})y_{i0}\} = 0.$$

For period $t = 4$, we have three moment conditions and three valid instruments:

$$E\{(u_{i4} - u_{i3})y_{i0}\} = 0$$
$$E\{(u_{i4} - u_{i3})y_{i1}\} = 0$$
$$E\{(u_{i4} - u_{i3})y_{i2}\} = 0.$$

All these moment conditions can be exploited in a GMM framework. To introduce the GMM estimator, define for general sample size T

$$\Delta\varepsilon_i = \begin{pmatrix} u_{i2} - u_{i1} \\ \cdots \\ u_{i,T} - u_{i,T-1} \end{pmatrix} \tag{10.53}$$

as the vector of transformed error terms, and

$$Z_i = \begin{pmatrix} [y_{i0}] & 0 & \cdots & 0 \\ 0 & [y_{i0}, y_{i1}] & & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & [y_{i0}, \ldots, y_{i,T-2}] \end{pmatrix} \tag{10.54}$$

as the matrix of instruments. Each row in the matrix Z_i contains the instruments that are valid for a given period. Consequently, the set of all moment conditions can be written concisely as

$$E\{Z_i'\Delta u_i\} = 0. \tag{10.55}$$

Note that these are $1 + 2 + 3 + \cdots + T - 1$ conditions. To derive the GMM estimator, write this as

$$E\{Z_i'(\Delta y_i - \gamma\Delta y_{i,-1})\} = 0. \tag{10.56}$$

Because the number of moment conditions will typically exceed the number of unknown coefficients, we estimate γ by minimizing a quadratic expression in terms of the corresponding sample moments (compare Chapter 5), that is,

$$\min_{\gamma} \left[\frac{1}{N}\sum_{i=1}^{N} Z_i'(\Delta y_i - \gamma\Delta y_{i,-1})\right]' W_N \left[\frac{1}{N}\sum_{i=1}^{N} Z_i'(\Delta y_i - \gamma\Delta y_{i,-1})\right], \tag{10.57}$$

where W_N is a symmetric positive definite weighting matrix.[14] Differentiating this with respect to γ and solving for γ gives

$$\hat{\gamma}_{GMM} = \left(\left(\sum_{i=1}^{N}\Delta y_{i,-1}'Z_i\right) W_N \left(\sum_{i=1}^{N} Z_i'\Delta y_{i,-1}\right)\right)^{-1} \times \left(\sum_{i=1}^{N}\Delta y_{i,-1}'Z_i\right) W_N \left(\sum_{i=1}^{N} Z_i'\Delta y_i\right). \tag{10.58}$$

The properties of this estimator depend upon the choice for W_N, although it is consistent as long as W_N is positive definite, for example, for $W_N = I$, the identity matrix.

The **optimal weighting matrix** is the one that gives the most efficient estimator, i.e. that gives the smallest asymptotic covariance matrix for $\hat{\gamma}_{GMM}$. From the general theory of GMM in Chapter 5, we know that the optimal weighting matrix is (asymptotically) proportional to the inverse of the covariance matrix of the sample moments. In this case, this means that the optimal weighting matrix should satisfy

$$\operatorname*{plim}_{N\to\infty} W_N = V\{Z_i'\Delta u_i\}^{-1} = E\{Z_i'\Delta u_i\Delta u_i'Z_i\}^{-1}. \tag{10.59}$$

In the standard case where no restrictions are imposed upon the covariance matrix of u_i, this can be estimated using a first-step consistent estimator of γ and replacing the expectation operator with a sample average. This gives

$$\hat{W}_N^{opt} = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\Delta\hat{u}_i\Delta\hat{u}_i'Z_i\right)^{-1}, \tag{10.60}$$

where $\Delta\hat{u}_i$ is a residual vector from a first-step consistent estimator, for example using $W_N = I$.

The general GMM approach does not impose that u_{it} is i.i.d. over individuals and time, and the optimal weighting matrix is thus estimated without imposing these restrictions. Note, however, that the absence of autocorrelation was needed to guarantee the validity of the moment conditions. Instead of estimating the optimal weighting matrix unrestrictedly, it is also possible (and potentially advisable in small samples) to impose the absence of autocorrelation in u_{it}, combined with a homoskedasticity assumption. Noting that under these restrictions

$$E\{\Delta u_i\Delta u_i'\} = \sigma_u^2 G = \sigma_u^2 \begin{pmatrix} 2 & -1 & 0 & \cdots \\ -1 & 2 & \ddots & 0 \\ 0 & \ddots & \ddots & -1 \\ \vdots & 0 & -1 & 2 \end{pmatrix}, \tag{10.61}$$

[14] The suffix N reflects that W_N can depend upon the sample size N and does not reflect the dimension of the matrix.

the optimal weighting matrix can be determined as

$$W_N^{opt} = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'GZ_i\right)^{-1}. \quad (10.62)$$

Note that this matrix does not involve unknown parameters, so that the optimal GMM estimator can be computed in one step if the original errors u_{it} are assumed to be homoskedastic and exhibit no autocorrelation.

Under weak regularity conditions, the GMM estimator for γ is asymptotically normal for $N \to \infty$ and fixed T, with its covariance matrix given by

$$\underset{N\to\infty}{\text{plim}} \left(\left(\frac{1}{N}\sum_{i=1}^{N}\Delta y_{i,-1}'Z_i\right)\left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\Delta u_i \Delta u_i' Z_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\Delta y_{i,-1}\right)\right)^{-1}. \quad (10.63)$$

This follows from the more general expressions in Section 5.6. With i.i.d. errors the middle term reduces to

$$\sigma_u^2 W_N^{opt} = \sigma_u^2\left(\frac{1}{N}\sum_{i=1}^{N} Z_i'GZ_i\right)^{-1}.$$

Alvarez and Arellano (2003) show that, in general, the GMM estimator is also consistent when both N and T tend to infinity, despite the fact that the number of moment conditions tends to infinity with the sample size. For large T, however, the GMM estimator will be close to the fixed effects estimator, which provides a more attractive alternative. Moreover, Windmeijer (2005) and others warn against using too many instruments in this context.

10.4.2 Dynamic Models with Exogenous Variables

If the model also contains exogenous variables, we have

$$y_{it} = x_{it}'\beta + \gamma y_{i,t-1} + \alpha_i + u_{it}, \quad (10.64)$$

which can also be estimated by the generalized instrumental variables or GMM approach. Depending upon the assumptions made about x_{it}, different sets of additional instruments can be constructed. If the x_{it} are *strictly exogenous* in the sense that they are uncorrelated with any of the u_{is} error terms, we also have

$$E\{x_{is}\Delta u_{it}\} = 0 \quad \text{for each } s, t, \quad (10.65)$$

so that $x_{i1}, \ldots, x_{iT}$ can be added to the instruments list for the first-differenced equation in each period. This would make the number of rows in Z_i quite large. Instead, almost the same level of information may be retained when the first-differenced x_{it}s are used as their own instruments.[15] In this case, we impose the moment conditions

$$E\{\Delta x_{it}\Delta u_{it}\} = 0 \quad \text{for each } t \quad (10.66)$$

[15] We give up potential efficiency gains if some x_{it} variables help 'explaining' the lagged endogenous variables.

and the instrument matrix can be written as

$$Z_i = \begin{pmatrix} [y_{i0}, \Delta x_{i2}'] & 0 & \cdots & 0 \\ 0 & [y_{i0}, y_{i1}, \Delta x_{i3}'] & & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & [y_{i0}, \ldots, y_{i,T-2}, \Delta x_{iT}'] \end{pmatrix}.$$

If the x_{it} variables are not strictly exogenous but **predetermined**, in which case current and lagged x_{it}s are uncorrelated with current error terms, we only have $E\{x_{it}u_{is}\} = 0$ for $s \geq t$. In this case, only $x_{i,t-1}, \ldots, x_{i1}$ are valid instruments for the first-differenced equation in period t. Thus, the moment conditions that can be imposed are

$$E\{x_{i,t-j}\Delta u_{it}\} = 0 \quad \text{for } j = 1, \ldots, t-1 \text{ (for each } t). \quad (10.67)$$

In practice, a combination of strictly exogenous and predetermined x variables may occur rather than one of these two extreme cases. The matrix Z_i should then be adjusted accordingly. Baltagi (2005, Chapter 8) provides additional discussion and examples.

Arellano and Bover (1995) provide a framework to integrate the above approach with the instrumental variables estimators of Hausman and Taylor (1981) and others discussed in Subsection 10.2.6. Most importantly, they discuss how information in levels can also be exploited in estimation. That is, in addition to the moment conditions presented above, it is also possible to exploit the presence of valid instruments for the levels equation (10.64) or its average over time (the between regression). This is of particular importance when the individual series are highly persistent. In this case, the GMM estimator may suffer from severe finite sample biases because the instruments are weak (see Subsection 5.5.4); see also Blundell and Bond (1998), Blundell, Bond and Windmeijer (2000) and Arellano (2003, Section 6.6). Under certain assumptions, suitably lagged differences of y_{it} can be used to instrument the equation in levels, in addition to the instruments for the first-differenced equation. For example, if $E\{\Delta y_{i,t-1}\alpha_i\} = 0$, $\Delta y_{i,t-1}$ can be used to instrument $y_{i,t-1}$ in (10.42) and

$$E\{(y_{it} - \gamma y_{i,t-1})(y_{i,t-1} - y_{i,t-2})\} = 0$$

is a valid moment condition that can be added (in the absence of serial correlation in u_{it}). The validity of this instrument depends upon the assumption that changes in y_{it} are uncorrelated with the fixed effects. This means that individuals are in a kind of steady state, in the sense that deviations from long-term values, conditional upon the exogenous variables, are not systematically related to α_i.

10.5 Illustration: Explaining Capital Structure

The capital structure of a firm tells us how a firm finances its operations, the most important sources being debt and equity. In their seminal paper, Modigliani and Miller (1958) show that in a frictionless world with efficient capital markets a firm's capital structure is irrelevant for its value. In reality, however, market imperfections, like taxes and bankruptcy costs, may make firm value depend on capital structure, and it can be argued that firms select optimal target debt ratios on the basis of a trade-off between the costs and benefits of debt. For example, firms would make a trade-off between the tax benefits of debt financing[16] and the costs of financial distress when they have borrowed too much. In this section, we follow Flannery and Rangan (2006) and investigate the explanatory power of the trade-off theory taking into account that firms may adjust only partially towards their target capital structure. This leads to a dynamic panel data model for the firm's debt ratio.

A firm's debt ratio measures the portion of a firm's capitalization financed with debt and can be defined as

$$MDR_{it} = \frac{D_{it}}{D_{it} + S_{it}P_{it}},$$

where D_{it} is the book value of a firm's interest-bearing debt, S_{it} is the number of common shares outstanding and P_{it} denotes the price per share, all at time t. If a firm is financed by a relatively great deal of debt, it is said to be highly leveraged. The optimal or target debt ratio of a firm at time t is assumed to depend upon firm characteristics, known at time $t-1$ and related to the costs and benefits of operating with various leverage ratios. Accordingly, the target debt ratio is assumed to satisfy

$$MDR^*_{it} = x'_{i,t-1}\beta + \eta_{it},$$

where η_{it} is a mean zero error term accounting for unobserved heterogeneity.

Adjustment costs may prevent firms from choosing their target debt ratio at each point in time. To accommodate this, we specify a target adjustment model as

$$MDR_{it} - MDR_{i,t-1} = (1-\gamma)(MDR^*_{it} - MDR_{i,t-1}),$$

where $0 \leq \gamma \leq 1$ (compare (9.10)). The coefficient γ measures the adjustment speed and is assumed to be identical across firms. If $\gamma = 0$, firms adjust immediately and completely to their target debt ratio. Combining the previous two equations, we can write

$$MDR_{it} = \gamma MDR_{i,t-1} + x'_{i,t-1}\beta(1-\gamma) + \varepsilon_{it},$$

where $\varepsilon_{it} = (1-\gamma)\eta_{it}$. Because it is likely that time-invariant unobserved firm-specific heterogeneity plays a role, our final specification is written as

$$MDR_{it} = \gamma MDR_{i,t-1} + x'_{i,t-1}\beta^* + \alpha_i + u_{it}, \quad (10.68)$$

which corresponds to a standard dynamic panel data model as discussed in the previous section.

The data we use and the choice of explanatory variables are similar to those in Flannery and Rangan (2006). Our sample of firms is taken from the Compustat Industrial Annual Tapes and covers the years 1987 to 2001 ($T = 15$), where we exclude financial firms and regulated utilities whose financing decisions may reflect special factors. Our final sample contains a random subsample of the larger panel covering $N = 3777$ firms

[16] In most countries interest payments are tax deductible.

and 19 573 firm-year observations.[17] The panel is unbalanced, with the average firm being observed for 5.2 years. To model the target debt ratio, the following variables are used:

ebit_ta	earnings before interest payments and taxes, divided by total assets
mb	ratio of market value to book value of assets
dep_ta	depreciation expenses as a proportion of fixed assets
log *(ta)*	log of total assets
fa_ta	proportion of fixed assets
rd_ta	research and development expenditures, divided by total assets (0 if missing)
rd_dum	dummy indicating whether *rd_ta* is missing
indmedian	industry median debt ratio
rated	dummy indicating whether the firm has a public debt rating

Because information on R&D expenditures is missing for a substantial proportion of the firm-years, we follow Flannery and Rangan (2006)'s pragmatic solution to add a dummy variable to the model equal to one if R&D information is missing. We first estimate the dynamic model in (10.68) by three estimators that are known to be inconsistent for $N \to \infty$ and fixed T: OLS, the within estimator from Subsection 10.2.1 and the first-difference estimator from Subsection 10.2.2. The results are presented in Table 10.3, where all standard errors are calculated in the panel-robust way. That is, standard errors are adjusted for heteroskedasticity and arbitrary forms of within-firm serial correlation (see Subsection 10.2.6). From Subsection 10.4.1, we expect that the OLS estimator for γ overestimates the true coefficient on the lagged dependent variable, while the within (fixed effects) estimator will underestimate it (see also Bond, 2002). The first-difference estimator is expected substantially to underestimate the true impact of the lagged dependent variable, particularly if γ is large. This can be understood from (10.45), noting that the first-difference estimator and the within estimator are identical for $T = 2$. These expectations are confirmed in Table 10.3.

The differences between the OLS, within and first-difference results are substantial. The OLS coefficient on lagged *MDR* of 0.883 implies that firms close only 11.7% of the gap between the current and target debt ratio within 1 year. This slow adjustment is consistent with the hypothesis that other considerations outweigh the cost of deviation from optimal leverage. However, the fixed effects approach estimates adjustment to be much faster, with an estimated adjustment speed of 46.5%. The first-difference estimate of −0.114 is simply ridiculous and is mainly presented here to show that inappropriate estimation techniques may yield strongly conflicting and economically senseless results. Given that the OLS and within estimates are probably biased in the opposite direction, we would expect the true adjustment speed to be between 0.535 and 0.884 (ignoring sampling error). Another notable difference between the columns in Table 10.3 is the estimated impact of firm size. The OLS estimate is statistically insignificant, while the within and first-difference estimates both yield a highly significant positive coefficient ($t = 12.39$ and $t = 12.61$ respectively). The latter results seem to make more sense, because large firms tend to operate with more leverage, for example because they have

[17] The data for this illustration are available as DEBTRATIO.

Table 10.3 OLS, within and OLS-FD estimation results dynamic model (panel-robust standard errors in parentheses)

Variable	OLS	within	first-difference
MDR_{t-1}	0.884	0.535	−0.114
	(0.005)	(0.012)	(0.012)
ebit_ta	−0.032	−0.050	−0.045
	(0.007)	(0.011)	(0.010)
mb	0.0016	0.0023	0.0028
	(0.0007)	(0.0010)	(0.0011)
dep_ta	−0.261	−0.124	0.110
	(0.035)	(0.071)	(0.079)
log(*ta*)	−0.0007	0.038	0.064
	(0.0006)	(0.003)	(0.005)
fa_ta	0.020	0.059	0.106
	(0.006)	(0.017)	(0.018)
rd_dum	0.007	0.0001	−0.017
	(0.002)	(0.0081)	(0.009)
rd_ta	−0.120	−0.066	−0.059
	(0.013)	(0.026)	(0.029)
indmedian	0.032	0.167	0.182
	(0.010)	(0.022)	(0.026)
rated	0.007	0.021	0.009
	(0.003)	(0.006)	(0.007)
within R^2		0.340	
between R^2		0.641	
overall R^2	0.741	0.563	0.028

better access to public debt markets. The industry median is included to control for industry characteristics that are not captured by the other explanatory variables and is expected to have a positive coefficient. The magnitude of the coefficient for *indmedian* is larger for the within and first-difference results than for OLS, and so is its statistical significance. The variable *rated* is potentially endogenous, as a firm's credit rating may depend upon its capital structure. We follow Flannery and Rangan (2006) and simply include *rated* as additional explanatory variable, noting that its inclusion or exclusion has little impact on the other coefficient estimates. Note that for most coefficients the OLS robust-standard errors are smaller than the within and first-difference ones. This makes sense as the latter two approaches allow for fixed effects and only identify the coefficients from the within variation in the data. For example, *rd_dum* exhibits very little time variation and therefore its effect is not very accurately estimated with the fixed effects approaches.

As mentioned before, all estimators in Table 10.3 are inconsistent. The first-difference estimator, while allowing for correlation between α_i and the explanatory variables, is severely biased because the first-differenced lagged dependent variable is highly negatively correlated with the first-differenced error term. The OLS results are inconsistent because of the correlation between the lagged debt ratio and α_i. Both biases do not disappear for $T \to \infty$. The within estimates also allow for fixed effects and thus for correlation between the unobservables in α_i and the explanatory variables, but they suffer from a small-T bias. Despite this, the latter results appear to make more sense

than the OLS ones, suggesting that controlling for firm-specific fixed effects in the target debt ratio is important.

To estimate the current dynamic panel data model consistently for $N \to \infty$ and fixed T, the Anderson–Hsiao instrumental variables estimators and the Arellano–Bond GMM estimators are potential candidates. Table 10.4 presents the estimation results of the different approaches. All estimators presented in this table are based on exploiting instruments for the first-differenced equation. The first column presents the results for the Anderson–Hsiao estimator when $\Delta MDR_{i,t-2}$ is used as an instrument for $\Delta MDR_{i,t-1}$, while the second column presents the results when the level $MDR_{i,t-2}$ is used to instrument $\Delta MDR_{i,t-1}$. The differences between the two columns are striking. The estimator using the first-differenced instrument suffers from very high standard errors and extremely unrealistic parameter estimates. For example, the estimated value for γ is as high as 7.03 with a (panel-robust) standard error of 7.32. The estimator using the level instrument seems to produce a bit more realistic results, although the estimated coefficient on the lagged dependent variable is larger than one. A potential explanation for the poor performance of the first-difference Anderson–Hsiao estimator is a weak instrument problem.[18] We can easily check this by inspecting the underlying reduced-form equations (compare Subsection 5.5.4). In a regression explaining $\Delta MDR_{i,t-1}$ from the first-differenced variables $\Delta x_{i,t-1}$ as well as the proposed instrument $\Delta MDR_{i,t-2}$, the panel-robust t-value of the latter variable is only −1.00. This suggest that the instrument $\Delta MDR_{i,t-2}$ is basically irrelevant and we should not take the corresponding results seriously. For the reduced form containing the instrument $MDR_{i,t-2}$, the corresponding t-value is −14.15. Although this indicates that the Anderson–Hsiao results using the level instrument do not suffer from a weak instrument problem, they yield an economically unappealing estimate of 1.358 for the lagged dependent variable. A potential explanation for this outcome is that the exogeneity of the instrument $MDR_{i,t-2}$ is violated because of the presence of serial correlation in u_{it}.

An alternative approach is the use of the Arellano and Bond (1991) estimator, where further lags of *MDR* are used as instruments for lagged *MDR* (in the first-differenced equation). The results of this are also presented in Table 10.4, where we assume that the explanatory variables are strictly exogenous. The one-step estimates are based on the optimal weighting matrix under the assumption of homoskedasticity given in (10.62), while the two-step estimates use the more generally valid weighting matrix from (10.60). Although several studies have reported that the two-step (optimal GMM) standard errors are biased downwards in small samples and recommend using the one-step estimates, in the current application they appear to be larger than the one-step ones. The one-step GMM results correspond to an adjustment speed of 25.1%, while the two-step estimates imply an annual adjustment of 22.8%. Overall, the standard errors of the GMM estimates are relatively high, and a substantial number of explanatory variables are individually statistically insignificant. Further, the GMM results suffer from two additional problems. First, the Sargan test of overidentifying restrictions based on the one-step estimates produces a highly significant test statistic of 781.20. Note, however,

[18] An alternative interpretation to this problem is given by Arellano (1989), who shows that with an autoregressive exogenous variable the Anderson–Hsiao estimator that uses first-differenced instruments has a singularity point and very large variances over a wide range of parameter values. The estimator that uses instruments in levels does not suffer from this problem.

Table 10.4 IV and GMM estimation results dynamic model

Variable	Anderson–Hsiao IV robust s.e.	robust s.e.	Arellano–Bond GMM one-step	two-step
MDR_{t-1}	7.033	1.358	0.749	0.772
	(7.325)	(0.091)	(0.032)	(0.036)
ebit_ta	1.208	0.203	0.099	0.098
	(1.305)	(0.026)	(0.012)	(0.015)
mb	0.244	0.047	0.029	0.026
	(0.247)	(0.004)	(0.002)	(0.002)
dep_ta	−1.858	−0.227	−0.066	−0.003
	(2.116)	(0.151)	(0.087)	(0.106)
log(*ta*)	−0.521	−0.053	0.005	0.003
	(0.607)	(0.013)	(0.005)	(0.007)
fa_ta	−1.091	−0.166	−0.062	−0.052
	(1.238)	(0.039)	(0.021)	(0.025)
rd_dum	−0.023	−0.021	−0.0178	−0.017
	(0.079)	(0.016)	(0.0100)	(0.11)
rd_ta	0.882	0.127	0.064	0.055
	(1.038)	(0.050)	(0.037)	(0.035)
indmedian	−3.378	−0.584	−0.061	−0.095
	(3.668)	(0.061)	(0.034)	(0.032)
rated	−0.272	−0.052	−0.021	−0.029
	(0.294)	(0.012)	(0.008)	(0.008)
Overidentifying restrictions test ($df = 104$)			887.17 ($p = 0.0000$)	437.11 ($p = 0.0000$)
Test for second-order autocorrelation in Δu_{it}			−3.39 ($p = 0.0007$)	−2.73 ($p = 0.0063$)
instruments:	ΔMDR_{t-2}	MDR_{t-2}	$MDR_{t-2}, MDR_{t-3}, \ldots$ (for each t)	

that this test is only valid under homoskedasticity. The two-step estimates produce a lower value for the test of overidentifying restrictions, but still highly significant. Second, the hypothesis of no serial correlation in u_{it}, which is required for the instruments to be valid, is strongly rejected for both GMM estimators. In addition, some of the GMM estimates are counterintuitive. For example, the effect of the industry median is estimated to be negative.

In summary, none of the reported estimates for the dynamic model to explain firms' debt ratios is entirely convincing. The (inconsistent) OLS and within results from Table 10.3 suggest that the true γ coefficient should be in the range 0.535–0.884 (although this ignores the estimation error in both estimates). While GMM yields γ estimates around 0.75, the overidentifying restrictions tests reject both for the one-step and for the two-step results and the coefficient estimates for several other variables are economically unappealing.

It should be noted here that, if the true coefficient on the lagged dependent variable is close to unity, lagged levels as employed in the Arellano–Bond procedure are poor instruments for first differences. Arellano and Bover (1995) and Blundell and Bond (1998) develop alternative estimators that are based on adding the original equation in levels to the system and using suitably lagged first differences as instruments. Obviously, these first differences should then be orthogonal to α_i.

10.6 Nonstationarity, Unit Roots and Cointegration

The recent literature exhibits an increasing integration of techniques and ideas from time series analysis, such as unit roots and cointegration, into the area of panel data modelling. The underlying reason for this development is that researchers have increasingly realized that cross-sectional information is a useful additional source of information that should be exploited. To analyse the effect of a certain policy measure, for example adopting a road tax or a pollution tax, it may be more fruitful to compare with other countries than to try to extract information about these effects from the country's own history. Pooling data from different countries may also help to overcome the problem that sample sizes of time series are fairly small, so that tests regarding long-run properties are not very powerful.

A number of recent articles discuss issues relating to unit roots, spurious regressions and cointegration in panel data. Most of this literature focuses upon the case in which the number of time periods T is fairly large, while the number of cross-sectional units N is small or moderate. As a consequence, it is quite important to deal with potential nonstationarity of the data series, while the presence of a unit root or cointegration may be of specific economic interest. For example, a wide range of applications exist concerning purchasing power parity, including Oh (1996), focusing on (non)stationarity of real exchange rates for a set of countries, or on testing for cointegration between nominal exchange rates and prices (compare Sections 8.5 and 9.3 and Subsection 9.5.4). For ease of discussion, we shall refer below to the cross-sectional units as countries, although they may also correspond to firms, industries or regions.

A crucial issue in analysing the time series on a number of countries simultaneously is that of heterogeneity. Because it is possible to estimate a separate regression for each country, it is natural to think of the possibility that model parameters are different across countries, a case commonly referred to as 'heterogeneous panels'. Robertson and Symons (1992) and Pesaran and Smith (1995) stress the importance of parameter heterogeneity in dynamic panel data models and analyse the potentially severe biases that may arise from handling it in an inappropriate manner; see also Canova (2007, Chapter 8). Such biases are particularly misleading in a nonstationary world as the relationships of the individual series may be completely destroyed.

As long as we consider each time series individually, and the series are of sufficient length, there is nothing wrong with applying the time series techniques from Chapters 8 and 9. However, if we pool different series, we have to be aware of the possibility that their processes do not all have the same characteristics or are not all described by the same parameters. For example, it is conceivable that y_{it} is stationary for country 1 but integrated of order one for country 2. Even when all variables are integrated of order one in each country, heterogeneity in cointegration properties may lead to problems. For example, if for each country i the variables y_{it} and x_{it} are cointegrated with parameter β_i, it holds that $y_{it} - \beta_i x_{it}$ is $I(0)$ for each i, but in general there does not exist a common cointegrating parameter β that makes $y_{it} - \beta x_{it}$ stationary for all i. Similarly, there is no guarantee that the cross-sectional averages $\bar{y}_t = (1/N)\sum_i y_{it}$ and $\bar{x}_t$ are cointegrated, even if all underlying individual series are cointegrated.

In Subsections 10.6.1 and 10.6.2, we pay attention to panel data unit root tests and cointegration tests respectively. Basically, the tests are directed at testing the joint null hypothesis of a unit root (or the absence of cointegration) for *each* of the countries involved. In comparison with the single time series case, panel data tests raise a number of additional issues, including cross-sectional dependence, heterogeneity in dynamics and error-term properties and the type of asymptotics that is employed. While most asymptotic analysis is done with both N and T tending to infinity, there are various ways that this can be done.

10.6.1 Panel Data Unit Root Tests

To introduce panel data unit root tests, consider the autoregressive model

$$y_{it} = \alpha_i + \gamma_i y_{i,t-1} + u_{it}, \tag{10.68}$$

which we can rewrite as

$$\Delta y_{it} = \alpha_i + \pi_i y_{i,t-1} + u_{it}, \tag{10.69}$$

where $\pi_i = \gamma_i - 1$. The null hypothesis that all series have a unit root then becomes H_0: $\pi_i = 0$ for all i. A first choice for the alternative hypothesis is that all series are stationary with the same mean-reversion parameter, that is, H_1: $\pi_i = \pi < 0$ for each country i, and is used in the approaches of Levin and Lin (1993),[19] Quah (1994), Harris and Tzavalis (1999) and Breitung (2000). A more general alternative allows the mean-reversion parameters to be potentially different across countries and states that H_1: $\pi_i < 0$ for at least one country i. This alternative is used by Maddala and Wu (1999), Choi (2001), Im, Pesaran and Shin (2003)[20] and others. As in the time series case discussed in Chapter 8, the properties of the test statistics (and their computation) depend crucially upon the deterministic regressors included in the test equation. For example, in (10.69) we have included a dummy for each country, corresponding to the fixed effect. Alternative tests are available in cases where the equation includes a common intercept, or in cases where a deterministic trend is added to the fixed effect.

For all tests, the null hypothesis is that the time series of *all* individual countries have a unit root. This implies that the null hypothesis can be rejected (in sufficiently large samples) if any one of the N coefficients π_i is less than zero. Rejection of the null hypothesis therefore does not indicate that all series are stationary. As Smith and Fuertes (2007) note, if the hypothesis of interest is that all series are stationary (for example, real exchange rates under purchasing power parity), it would be more appropriate to use a panel version of the KPSS test, as discussed in Section 8.4, where stationarity is the null hypothesis rather than the alternative; see Hadri (2000). However, a test like this may reject if just one series is nonstationary, which may not be interesting either. Because of these issues, Maddala, Wu and Liu (2000) argue that, for purchasing power parity, panel data unit root tests are the wrong answer to the low power of unit root tests in single time series.

In addition to the choice of deterministic regressors in the test equations, panel data unit root tests offer three additional technical issues in comparison with the single time series case. First, one has to make assumptions on the cross-sectional dependence between u_{it}s, noting that a majority of the existing nonstationary panel data literature assumes cross-sectional independence. Second, we need to be specific on the properties

[19] A revised version of the Levin and Lin (1993) paper is available in Levin, Lin and Chu (2002).
[20] A first version of this paper dates back to 1995.

of u_{it} and how they are allowed to vary across the different units. This includes serial correlation and the possibility of heteroskedasticity across units. Third, asymptotic properties of estimators and tests depend crucially upon the way in which N, the number of cross-sectional units, and T, the number of time periods, tend to infinity (see Phillips and Moon, 1999). Some tests assume that either T or N is fixed and assume that the other dimension tends to infinity. Many tests are based on a sequential limit, where first T tends to infinity for fixed N, and subsequently N tends to infinity. Alternatively, some tests assume that both N and T tend to infinity along a specific path (e.g. T/N being fixed). While the type of asymptotics that is applied may seem a theoretical issue, remember that we are using asymptotic theory to approximate the properties of estimators and tests in the finite sample that we happen to have. Although it is hard to make general statements on this matter, some asymptotic approximations are simply better than others. Many papers in this area therefore also contain a Monte Carlo study to analyse the finite sample behaviour of the proposed tests under controlled circumstances. A common finding for many of the tests below is that they tend to be oversized. That is, when the null hypothesis is true, the tests tend to reject more frequently than their nominal size (say, 5%) suggests. Further, many tests do not perform very well when the error terms are cross-sectionally correlated, or in the presence of cross-country cointegration. For example, when real exchange rates are $I(1)$ and cointegrated across countries, the null hypothesis tends to be rejected too often (see Banerjee, Marcellino and Osbat, 2005, for an illustration).

While it is beyond the scope of this text to discuss alternative panel data unit root tests in great technical detail, a brief discussion of some tests is warranted. More details can be found in Banerjee (1999), Enders (2004, Section 4.11) or Baltagi (2005, Chapter 12). Levin and Lin (1993) and Harris and Tzavalis (1999) base their tests upon the OLS estimator for π, assuming that u_{it} is i.i.d. across countries and time. Depending upon the deterministic regressors included, the OLS estimator may be biased, even asymptotically. When fixed effects are included, the estimator corresponds to the fixed effects estimator for π based on (10.69), which is biased for fixed T (see Section 10.4). With appropriate correction and standardization factors, test statistics can be derived that are asymptotically normal for $N \to \infty$ and fixed T (Harris and Tzavalis, 1999) or both $N, T \to \infty$ (Levin and Lin, 1993); see Baltagi (2005, Section 12.2). While the test statistics can be modified to allow for serial correlation in u_{it}, they do not allow cross-sectional dependence. This assumption is rather strong, and, as stressed by O'Connell (1998) in a panel study on purchasing power parity, allowing for cross-sectional dependence may substantially affect inferences about the presence of a unit root. Because individual observations in a panel typically have no natural ordering, modelling cross-sectional dependence is not obvious.

The above two sets of tests are restrictive because they assume that π_i is the same across all countries, also under the alternative hypothesis. The test proposed by Im, Pesaran and Shin (2003) allows π_i to be different across individual units. It is based on averaging the augmented Dickey–Fuller (ADF) test statistics (see Section 8.4) over the cross-sectional units, while allowing for different orders of serial correlation. In fact, the alternative hypothesis states that $\pi_i < 0$ for at least one i and thus allows that $\pi_i = 0$ for a subset of the countries. Im, Pesaran and Shin (2003) also propose a test based on the N Lagrange multiplier statistics for $\pi_i = 0$, averaged over all countries. The idea underlying these tests is quite simple: if you have N independent test statistics, their average will be asymptotically normally distributed for $N \to \infty$. Consequently, the tests are based on comparison of appropriately scaled cross-sectional averages with critical values from a standard normal distribution.

An alternative approach to combining information from individual unit root tests is employed by Maddala and Wu (1999) and Choi (2001), who propose panel data unit root tests based on combining the p-values of the N cross-sectional tests. Let p_i denote the p-value of the (augmented) Dickey–Fuller test for unit i. Under the null hypothesis, p_i will have a uniform distribution over the interval [0, 1], small values corresponding to rejection. The combined test statistic is given by

$$P = -2\sum_{i=1}^{N} \log p_i. \tag{10.70}$$

For fixed N, this test statistic will have a Chi-squared distribution with $2N$ degrees of freedom as $T \to \infty$, so that large values of P lead us to reject the null hypothesis. While this test (sometimes referred to as the Fisher test) is attractive because it allows the use of different ADF tests and different time-series lengths per unit, a disadvantage is that it requires individual p-values that have to be derived by Monte Carlo simulations.

While the latter tests may seem attractive and easy to use, a word of caution is appropriate. Before one can apply the individual ADF tests underlying the Maddala and Wu (1999) and Im, Pesaran and Shin (2003) approaches, one has to determine the number of lags and determine whether a trend should be included. It is not obvious how this should be done. For a single time series, a common approach is to perform the ADF test for a range of alternative lag values. For example, in Table 8.2 we presented 26 different (augmented) Dickey–Fuller test statistics for the log price index. If we were to combine the ADF tests for N different countries, in whatever way, this would create a wide range of possible combinations. Smith and Fuertes (2007) warn for pretest biases in this context.

10.6.2 Panel Data Cointegration Tests

A wide range of alternative tests is available to test for cointegration in a dynamic panel data setting, and research in this area is evolving rapidly. A substantial number of these tests are based on testing for a unit root in the residuals of a panel cointegrating regression. The drawbacks and complexities associated with the panel unit root tests are also relevant in the cointegration case. Several additional issues are of potential importance when testing for cointegration: heterogeneity in the parameters of the cointegrating relationships, heterogeneity in the number of cointegrating relationships across countries and the possibility of cointegration between the series from different countries. A final issue is that of estimating the cointegrating vectors, for which several alternative estimators are available, with different small- and large-sample properties (depending upon the type of asymptotics that is chosen).

When the cointegrating relationship is unknown, which is almost always the case, most cointegration tests start with estimating the cointegrating regression. Let us focus on the bivariate case and write the panel regression as

$$y_{it} = \alpha_i + \beta_i x_{it} + u_{it}, \tag{10.71}$$

where both y_{it} and x_{it} are integrated of order one. Cointegration implies that u_{it} is stationary for each i. Homogeneous cointegration, in addition, requires that $\beta_i = \beta$. If the cointegrating parameter is heterogeneous, and homogeneity is imposed, one estimates

$$y_{it} = \alpha_i + \beta x_{it} + [(\beta_i - \beta)x_{it} + u_{it}], \tag{10.72}$$

and in general the composite error term is integrated of order one, even if u_{it} is stationary. However, the problem of spurious regressions may be less relevant in this situation. This is because a pooled estimator will also average over i, so that the noise in the equation will be attenuated. In many circumstances, when $N \to \infty$, the fixed effects estimator for β is actually consistent for the long-run average relation parameter, as well as asymptotically normal, despite the absence of cointegration (see Phillips and Moon, 1999). However, the meaning of this long-run relationship, in the absence of cointegration, is open to some interpretation (see Hsiao, 2003, Section 10.2, for some discussion). With heterogeneous cointegration, the long-run average estimated from the pooled regression may differ substantially from the average of the cointegration parameters averaged over countries (see Pesaran and Smith, 1995). Consequently, if there is heterogeneous cointegration, it is much better to estimate the individual cointegrating regressions rather than using a pooled estimator. Obviously, this requires $T \to \infty$.

To test for cointegration, the panel data unit root tests from the previous section can be applied to the residuals from these regressions, provided that the critical values are appropriately adjusted (see Pedroni, 1999, or Kao, 1999). Recall that these tests assume cross-sectional independence. Some tests assume homogeneity of the cointegrating parameter and use a pooled OLS or dynamic OLS estimator (see Subsection 9.2.2). Additional discussion on these tests can be found in Banerjee (1999), Baltagi (2005, Section 12.5), Smith and Fuertes (2007) or Breitung and Pesaran (2008).

With more than two variables, an additional complication may arise because more than one cointegrating relationship may exist for one or more of the countries. Further, even with one cointegrating vector per country, the results will be sensitive to the normalization constraint (left-hand-side variable) that is chosen. Finally, the existence of between-country cointegration may seriously distort the results of within-country cointegration tests.

10.7 Models with Limited Dependent Variables

Panel data are relatively often used in micro-economic problems where the models of interest involve nonlinearities. Discrete or limited dependent variables are an important phenomenon in this area, and their combination with panel data usually complicates estimation. The reason is that with panel data it can usually not be argued that different observations on the same unit are independent. Correlations between different error terms typically complicate the likelihood functions of such models and therefore complicate their estimation. In this section we discuss the estimation of panel data logit, probit and tobit models. More details on panel data models with limited dependent variables can be found in Maddala (1987) or Hsiao (2003, Chapters 7–8).

10.7.1 Binary Choice Models

As in the cross-sectional case, the binary choice model is usually formulated in terms of an underlying latent model. Typically, we write[21]

$$y_{it}^* = x_{it}'\beta + \alpha_i + u_{it}, \tag{10.73}$$

where we observe $y_{it} = 1$ if $y_{it}^* > 0$ and $y_{it} = 0$ otherwise. For example, y_{it} may indicate whether person i is working in period t or not. Let us assume that the idiosyncratic error term u_{it} has a symmetric distribution with distribution function $F(.)$, i.i.d. across individuals and time and independent of all x_{is}. Even in this case the presence of α_i complicates estimation, both when we treat them as fixed unknown parameters and when we treat them as random error terms.

If we treat α_i as fixed unknown parameters, we are essentially including N dummy variables in the model. The loglikelihood function is thus given by (compare (7.12))

$$\log L(\beta, \alpha_1, \ldots, \alpha_N) = \sum_{i,t} y_{it} \log F(\alpha_i + x_{it}'\beta) + \sum_{i,t} (1 - y_{it}) \log[1 - F(\alpha_i + x_{it}'\beta)]. \tag{10.74}$$

Maximizing this with respect to β and α_i $(i = 1, \ldots, N)$ results in consistent estimators *provided that the number of time periods T goes to infinity*. For fixed T and $N \to \infty$, the estimators are inconsistent. The reason is that, for fixed T, the number of parameters grows with sample size N and we have what is known as an 'incidental parameter' problem. Clearly, we can only estimate α_i consistently if the number of observations for individual i grows, which requires that T tends to infinity. In general, the inconsistency of $\hat{\alpha}_i$ for fixed T will carry over to the estimator for β.

The incidental parameter problem, where the number of parameters increases with the number of observations, arises in any fixed effects model, including the linear model; see Lancaster (2000) for a recent discussion. For the linear case, however, it was possible to eliminate the α_is, such that β could be estimated consistently, even though all the α_i parameters could not. For most nonlinear models, however, the inconsistency of $\hat{\alpha}_i$ leads to inconsistency of the other parameter estimators as well. Also note that, from a practical point of view, the estimation of more than N parameters may not be very attractive if N is fairly large.

Although it is possible to transform the *latent* model such that the individual effects α_i are eliminated, this does not help in this context because there is no mapping from, for example, $y_{it}^* - y_{i,t-1}^*$ to observables like $y_{it} - y_{i,t-1}$. An alternative strategy is the use of **conditional maximum likelihood** (see Andersen, 1970, or Chamberlain, 1980). In this case, we consider the likelihood function conditional upon a set of statistics t_i that are sufficient for α_i. This means that, conditional upon t_i, an individual's likelihood contribution no longer depends upon α_i but still depends upon the other parameters β.

[21] To simplify the notation, we shall assume that x_{it} includes a constant, whenever appropriate.

In the panel data binary choice model, the existence of a sufficient statistic depends upon the functional form of F, that is, depends upon the distribution of u_{it}.

At the general level let us write the joint density or probability mass function of $y_{i1}, \ldots, y_{iT}$ as $f(y_{i1}, \ldots, y_{iT}|\alpha_i, \beta)$, which depends upon the parameters β and α_i. If a sufficient statistic t_i exists, this means that there exists an observable variable t_i such that $f(y_{i1}, \ldots, y_{iT}|t_i, \alpha_i, \beta) = f(y_{i1}, \ldots, y_{iT}|t_i, \beta)$ and so does not depend upon α_i. Consequently, we can maximize the **conditional likelihood function**, based upon $f(y_{i1}, \ldots, y_{iT}|t_i, \beta)$, to get a consistent estimator for β. Moreover, we can use all the distributional results from Chapter 6 if we replace the loglikelihood with the conditional loglikelihood function. For the *linear* model with normal errors, a sufficient statistic for α_i is $\bar{y}_i$. That is, the conditional distribution of y_{it} given $\bar{y}_i$ does not depend upon α_i, and maximizing the conditional likelihood function can be shown to reproduce the fixed effects estimator for β. Unfortunately, this result does not automatically extend to nonlinear models. For the probit model, for example, it has been shown that no sufficient statistic for α_i exists. This means that we cannot estimate a fixed effects probit model consistently for fixed T.

10.7.2 The Fixed Effects Logit Model

For the fixed effects logit model, the situation is different. In this model $t_i = \bar{y}_i$ is a sufficient statistic for α_i and consistent estimation is possible by conditional maximum likelihood. It should be noted that the conditional distribution of $y_{i1}, \ldots, y_{iT}$ is degenerate if $t_i = 0$ or $t_i = 1$. Consequently, such individuals do not contribute to the conditional likelihood and should be discarded in estimation. Put differently, their behaviour would be completely captured by their individual effect α_i. This means that only individuals that change status at least once are relevant for estimating β. To illustrate the fixed effects logit model, we consider the case with $T = 2$.

By conditioning upon $t_i = 1/2$, we restrict the sample to the observations for which y_{it} changes, and the two possible outcomes are (0, 1) and (1, 0). The conditional probability of the first outcome is

$$P\{(0,1)|t_i = 1/2, \alpha_i, \beta\} = \frac{P\{(0,1)|\alpha_i, \beta\}}{P\{(0,1)|\alpha_i, \beta\} + P\{(1,0)|\alpha_i, \beta\}}. \tag{10.75}$$

Using

$$P\{(0,1)|\alpha_i, \beta\} = P\{y_{i1} = 0|\alpha_i, \beta\} P\{y_{i2} = 1|\alpha_i, \beta\}$$

with[22]

$$P\{y_{i2} = 1|\alpha_i, \beta\} = \frac{\exp\{\alpha_i + x_{i2}'\beta\}}{1 + \exp\{\alpha_i + x_{i2}'\beta\}},$$

it follows that the conditional probability is given by

$$P\{(0,1)|t_i = 1/2, \alpha_i, \beta\} = \frac{\exp\{(x_{i2} - x_{i1})'\beta\}}{1 + \exp\{(x_{i2} - x_{i1})'\beta\}}, \tag{10.76}$$

[22] See (7.6) in Chapter 7 for the logistic distribution function.

which indeed does not depend upon α_i. Similarly,

$$P\{(1,0)|t_i = 1/2, \alpha_i, \beta\} = \frac{1}{1+\exp\{(x_{i2}-x_{i1})'\beta\}}. \qquad (10.77)$$

These results show that the conditional distribution of (y_{i1}, y_{i2}), given t_i and α_i, is independent of the individual specific effects. Accordingly, we can estimate the fixed effect logit model for $T = 2$ using a standard logit with $x_{i2} - x_{i1}$ as explanatory variables and the change in y_{it} as the endogenous event (1 for a positive change, 0 for a negative one). In a sense, conditioning upon $t_i = 1/2$ has the same effect as first differencing (or within transforming) the data in a linear panel data model. Note that in this fixed effects binary choice model it is even more clear than in the linear case that the model is only identified through the 'within dimension' of the data; individuals who do not change status are simply discarded in estimation as they provide no information whatsoever about β. For the case with larger T, it is a bit more cumbersome to derive all the necessary conditional probabilities, but in principle it is a straightforward extension of the above case (see Chamberlain, 1980, or Maddala, 1987). Chamberlain (1980) also discusses how the conditional maximum likelihood approach can be extended to the multinomial logit model.

If it can be assumed that the α_i are independent of the explanatory variables in x_{it}, a random effects treatment seems more appropriate. This is most easily achieved in the context of a probit model.

10.7.3 The Random Effects Probit Model

Let us start with the latent variable specification

$$y_{it}^* = x_{it}'\beta + \varepsilon_{it}, \qquad (10.78)$$

with

$$y_{it} = 1 \quad \text{if } y_{it}^* > 0$$
$$y_{it} = 0 \quad \text{if } y_{it}^* \leq 0, \qquad (10.79)$$

where ε_{it} is an error term with mean zero and unit variance, independent of $(x_{i1}, \ldots, x_{iT})$. To estimate β by maximum likelihood, we will have to complement this with an assumption about the joint distribution of $\varepsilon_{i1}, \ldots, \varepsilon_{iT}$. The likelihood contribution of individual i is the (joint) probability of observing the T outcomes $y_{i1}, \ldots, y_{iT}$. This joint probability is determined from the joint distribution of the latent variables $y_{i1}^*, \ldots, y_{iT}^*$ by integrating over the appropriate intervals. In general, this will thus imply T integrals, which in estimation are typically to be computed numerically. When $T = 4$ or more, this makes maximum likelihood estimation infeasible. It is possible to circumvent this 'curse of dimensionality' by using simulation-based estimators, as discussed in, for example, Keane (1993), Weeks (1995) and Hajivassiliou and McFadden (1998). Their discussion is beyond the scope of this text.

Clearly, if it can be assumed that all ε_{it} are independent, we have $f(y_{i1}, \ldots, y_{iT}|x_{i1}, \ldots, x_{iT}, \beta) = \prod_t f(y_{it}|x_{it}, \beta)$, which involves T one-dimensional integrals only (as in the cross-sectional case). If we make an error components assumption,

and assume that $\varepsilon_{it} = \alpha_i + u_{it}$, where u_{it} is independent over time (and individuals), we can write the joint probability as

$$f(y_{i1}, \ldots, y_{iT}|x_{i1}, \ldots, x_{iT}, \beta) = \int_{-\infty}^{\infty} f(y_{i1}, \ldots, y_{iT}|x_{i1}, \ldots, x_{iT}, \alpha_i, \beta) f(\alpha_i) d\alpha_i$$
$$= \int_{-\infty}^{\infty} \left[\prod_t f(y_{it}|x_{it}, \alpha_i, \beta)\right] f(\alpha_i) d\alpha_i, \qquad (10.80)$$

which requires numerical integration over one dimension. This is a feasible specification that allows the error terms to be correlated across different periods, albeit in a restrictive way. The crucial step in (10.80) is that, conditional upon α_i, the errors from different periods are independent.

In principle, arbitrary assumptions can be made about the distributions of α_i and u_{it}. For example, one could assume that u_{it} is i.i.d. normal while α_i has a logistic distribution. However, this may lead to distributions for $\alpha_i + u_{it}$ that are nonstandard. For example, the sum of two logistically distributed variables in general does not have a logistic distribution. This implies that individual probabilities, like $f(y_{it}|x_{it}, \beta)$, are hard to compute and do not correspond to a cross-sectional probit or logit model. Therefore, it is more common to start from the joint distribution of $\varepsilon_{i1}, \ldots, \varepsilon_{iT}$. The multivariate logistic distribution has the disadvantage that all correlations are restricted to be 1/2 (see Maddala, 1987), so that it is not very attractive in practice. Consequently, the most common approach is to start from a multivariate normal distribution, which leads to the **random effects probit model**.

Let us assume that the joint distribution of $\varepsilon_{i1}, \ldots, \varepsilon_{iT}$ is normal with zero means and variances equal to 1 and $\text{cov}\{\varepsilon_{it}, \varepsilon_{is}\} = \sigma_\alpha^2, s \neq t$. This corresponds to assuming that α_i is $NID(0, \sigma_\alpha^2)$ and u_{it} is $NID(0, 1 - \sigma_\alpha^2)$. Recall that, as in the cross-sectional case, we need a normalization on the errors' variances. The normalization chosen here implies that the error variance in a given period is unity, such that the estimated β coefficients are directly comparable with estimates obtained from estimating the model from one wave of the panel using cross-sectional probit maximum likelihood. For the random effects probit model, the expressions in the likelihood function are given by

$$f(y_{it}|x_{it}, \alpha_i, \beta) = \Phi\left(\frac{x_{it}'\beta + \alpha_i}{\sqrt{1-\sigma_\alpha^2}}\right) \quad \text{if } y_{it} = 1$$
$$= 1 - \Phi\left(\frac{x_{it}'\beta + \alpha_i}{\sqrt{1-\sigma_\alpha^2}}\right) \quad \text{if } y_{it} = 0, \qquad (10.81)$$

where Φ denotes the cumulative density function of the standard normal distribution. The density of α_i is given by

$$f(\alpha_i) = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left\{-\frac{1}{2}\frac{\alpha_i^2}{\sigma_\alpha^2}\right\}. \qquad (10.82)$$

The integral in (10.80) has to be computed numerically, which can be done using the algorithm described in Butler and Moffitt (1982). Several software packages (for example, LIMDEP and Stata) have standard routines for estimating the random effects probit model.

It can be shown (Robinson, 1982) that ignoring the correlations across periods and estimating the β coefficients using standard probit maximum likelihood on the pooled data is consistent, though inefficient. Moreover, routinely computed standard errors are incorrect. Nevertheless, these values can be used as initial estimates in an iterative maximum likelihood procedure based on (10.80).

10.7.4 Tobit Models

The random effects tobit model is very similar to the random effects probit model, the only difference being in the observation rule. Consequently, we can be fairly brief here. Let us start with

$$y_{it}^* = x_{it}'\beta + \alpha_i + u_{it}, \quad (10.83)$$

while

$$\begin{aligned} y_{it} &= y_{it}^* && \text{if } y_{it}^* > 0 \\ y_{it} &= 0 && \text{if } y_{it}^* \leq 0. \end{aligned} \quad (10.84)$$

We make the usual random effects assumption that α_i and u_{it} are i.i.d. normally distributed, independent of $x_{i1}, \ldots, x_{iT}$, with zero means and variances σ_α^2 and σ_u^2 respectively. Using f as generic notation for a density or probability mass function, the likelihood function can be written as in (10.80):

$$f(y_{i1}, \ldots, y_{iT}|x_{i1}, \ldots, x_{iT}, \beta) = \int_{-\infty}^{\infty} \prod_t f(y_{it}|x_{it}, \alpha_i, \beta) f(\alpha_i) d\alpha_i,$$

where $f(\alpha_i)$ is given by (10.82) and $f(y_{it}|x_{it}, \alpha_i, \beta)$ is given by

$$\begin{aligned} f(y_{it}|x_{it}, \alpha_i, \beta) &= \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left\{-\frac{1}{2}\frac{(y_{it} - x_{it}'\beta - \alpha_i)^2}{\sigma_u^2}\right\} && \text{if } y_{it} > 0 \\ &= 1 - \Phi\left(\frac{x_{it}'\beta + \alpha_i}{\sigma_u}\right) && \text{if } y_{it} = 0. \end{aligned} \quad (10.85)$$

Note that the latter two expressions are similar to the likelihood contributions in the cross-sectional case, as discussed in Chapter 7. The only difference is the inclusion of α_i in the conditional mean.

In a completely similar fashion, other forms of censoring can be considered, to obtain, for example, the random effects ordered probit model. In all cases, the integration over α_i has to be done numerically.

10.7.5 Dynamics and the Problem of Initial Conditions

The possibility of including a lagged dependent variable in the above models is of economic interest. For example, suppose we are explaining whether or not an individual is unemployed over a number of consecutive months. It is typically the case that

individuals who have a longer history of being unemployed are less likely to leave the state of unemployment. As discussed in the introductory section of this chapter, there are two explanations for this: an individual with a longer unemployment history may be discouraged in looking for a job or may (for whatever reason) be less attractive for an employer to hire. This is referred to as **state dependence**: the longer you are in a certain state, the less likely you are to leave it. Alternatively, it is possible that **unobserved heterogeneity** is present such that individuals with certain unobserved characteristics are less likely to leave unemployment. The fact that we observe a spurious state dependence in the data is simply due to a selection mechanism: the long-term unemployed have certain unobservable (time-invariant) characteristics that make it less likely for them to find a job anyhow. In the binary choice models discussed above, the individual effects α_i capture the unobserved heterogeneity. If we include a lagged dependent variable, we can distinguish between the above two explanations.

Let us consider the random effect probit model, although similar results hold for the random effects tobit case. Suppose the latent variable specification is changed into

$$y_{it}^* = x_{it}'\beta + \gamma y_{i,t-1} + \alpha_i + u_{it}, \quad (10.86)$$

with $y_{it} = 1$ if $y_{it}^* > 0$ and 0 otherwise. In this model $\gamma > 0$ indicates positive state dependence: the ceteris paribus probability that $y_{it} = 1$ is larger if $y_{i,t-1}$ is also one. Let us consider maximum likelihood estimation of this dynamic random effects probit model, making the same distributional assumptions as before. In general terms, the likelihood contribution of individual i is given by[23]

$$\begin{aligned} &f(y_{i1}, \ldots, y_{iT}|x_{i1}, \ldots, x_{iT}, \beta) \\ &= \int_{-\infty}^{\infty} f(y_{i1}, \ldots, y_{iT}|x_{i1}, \ldots, x_{iT}, \alpha_i, \beta) f(\alpha_i) d\alpha_i \\ &= \int_{-\infty}^{\infty} \left[\prod_{t=2}^{T} f(y_{it}|y_{i,t-1}, x_{it}, \alpha_i, \beta)\right] f(y_{i1}|x_{i1}, \alpha_i, \beta) f(\alpha_i) d\alpha_i, \end{aligned} \quad (10.87)$$

where

$$\begin{aligned} f(y_{it}|y_{i,t-1}, x_{it}, \alpha_i, \beta) &= \Phi\left(\frac{x_{it}'\beta + \gamma y_{i,t-1} + \alpha_i}{\sqrt{1 - \sigma_\alpha^2}}\right) && \text{if } y_{it} = 1 \\ &= 1 - \Phi\left(\frac{x_{it}'\beta + \gamma y_{i,t-1} + \alpha_i}{\sqrt{1 - \sigma_\alpha^2}}\right) && \text{if } y_{it} = 0. \end{aligned}$$

This is completely analogous to the static case, and $y_{i,t-1}$ is simply included as an additional explanatory variable. However, the term $f(y_{i1}|x_{i1}, \alpha_i, \beta)$ in the likelihood function may cause problems. It gives the probability of observing $y_{i1} = 1$ or 0 without knowing the previous state but conditional upon the unobserved heterogeneity term α_i.

If the initial value is exogenous in the sense that its distribution does not depend upon α_i, we can put the term $f(y_{i1}|x_{i1}, \alpha_i, \beta) = f(y_{i1}|x_{i1}, \beta)$ outside the integral.

[23] For notational convenience, the time index is defined such that the first observation is (y_{i1}, x_{i1}').

In this case, we can simply consider the likelihood function conditional upon y_{i1} and ignore the term $f(y_{i1}|x_{i1}, \beta)$ in estimation. The only consequence may be a loss of efficiency if $f(y_{i1}|x_{i1}, \beta)$ provides information about β. This approach would be appropriate if the initial state were necessarily the same for all individuals or if it were randomly assigned to individuals. An example of the first situation is given in Nijman and Verbeek (1992), who model nonresponse with respect to consumption. In their application the initial period refers to the month before the panel and no nonresponse was necessarily observed.

However, it may be hard to argue in many applications that the initial value y_{i1} is exogenous and does not depend upon a person's unobserved heterogeneity. In that case we would need an expression for $f(y_{i1}|x_{i1}, \alpha_i, \beta)$, and this is problematic. If the process that we are estimating has been going on for a number of periods before the current sample period, $f(y_{i1}|x_{i1}, \alpha_i, \beta)$ is a complicated function that depends upon person i's unobserved history. This means that it is typically impossible to derive an expression for the marginal probability $f(y_{i1}|x_{i1}, \alpha_i, \beta)$ that is consistent with the rest of the model. Heckman (1981) suggests an approximate solution to this **initial conditions problem** that appears to work reasonably well in practice. It requires an approximation for the marginal probability of the initial state by a probit function, using as much presample information as available, without imposing restrictions between its coefficients and the structural β and γ parameters. Hyslop (1999) employs this approach to estimate a dynamic model of female labour force participation; Vella and Verbeek (1999a) provide an illustration in the context of a dynamic random effects tobit model. The impact of the initial conditions diminishes if the number of sample periods T increases, so one may decide to ignore the problem when T is fairly large; see Hsiao (2003, Subsection 7.5.2) for more discussion.

10.7.6 *Semi-parametric Alternatives*

The binary choice and censored regression models discussed above suffer from two important drawbacks. First, the distribution of u_{it} conditional upon x_{it} (and α_i) needs to be specified, and second, with the exception of the fixed effects logit model, there is no simple way to estimate the models treating α_i as fixed unknown parameters. Several semi-parametric approaches have been suggested for these models that do not require strong distributional assumptions on u_{it} and somehow allow α_i to be eliminated before estimation.

In the binary choice model, it is possible to obtain semi-parametric estimators for β that are consistent up to a scaling factor whether or not α_i is treated as fixed or random. For example, Manski (1987) suggests a maximum score estimator (compare Subsection 7.1.8), while Lee (1999) provides a $\sqrt{N}$-consistent estimator for the static binary choice model; see Hsiao (2003, Section 7.4) for more discussion. Honoré and Kyriazidou (2000) propose a semi-parametric estimator for discrete choice models with a lagged dependent variable.

A tobit model as well as a truncated regression model with fixed effects can be estimated consistently using the generalized method of moments exploiting the moment conditions given by Honoré (1992) or Honoré (1993) for the dynamic model. The essential trick of these estimators is that a first-difference transformation, for appropriate subsets of the observations, no longer involves the incidental parameters α_i; see Hsiao (2003, Sections 8.4 and 8.6) for more discussion.

10.8 Incomplete Panels and Selection Bias

For a variety of reasons, empirical panel data sets are often incomplete. For example, after a few waves of the panel, people may refuse cooperation, households may not be located again or may have split up, firms may have finished business or may have merged with another firm or investment funds may be closed down. On the other hand, firms may enter business at a later stage, refreshment samples may have been drawn to compensate attrition or the panel may be collected as a rotating panel. In a rotating panel, each period a fixed proportion of the units is replaced. A consequence of all these events is that the resulting panel data set is no longer rectangular. If the total number of individuals equals N and the number of time periods is T, then the total number of observations is substantially smaller than NT.

A first consequence of working with an incomplete panel is a computational one. Most of the expressions for the estimators given above are no longer appropriate if observations are missing. A simple 'solution' is to discard any individual from the panel that has incomplete information and to work with the completely observed units only. In this approach, estimation uses the **balanced subpanel** only. This is computationally attractive but potentially highly inefficient: a substantial amount of information may be 'thrown away'. This loss in efficiency can be prevented by using all observations including those on individuals that are not observed in all T periods. This way, one uses the **unbalanced panel**. In principle this is straightforward, but computationally it requires some adjustments to the formulae in the previous sections. We shall discuss some of these adjustments in Subsection 10.8.1. Fortunately, most software that can handle panel data also allows for unbalanced data.

Another potential and even more serious consequence of using incomplete panel data is the danger of **selection bias**. If individuals are incompletely observed for an endogenous reason, the use of either the balanced subpanel or the unbalanced panel may lead to biased estimators and misleading tests. To elaborate upon this, suppose that the model of interest is given by

$$y_{it} = x_{it}'\beta + \alpha_i + u_{it}. \tag{10.88}$$

Furthermore, define the indicator variable r_{it} ('response') as $r_{it} = 1$ if (x_{it}, y_{it}) is observed and 0 otherwise. The observations on (x_{it}, y_{it}) are **missing at random** if r_{it} is independent of α_i and u_{it}. This means that conditioning upon the outcome of the selection process does not affect the conditional distribution of y_{it} given x_{it}. If we want to concentrate upon the balanced subpanel, the conditioning is upon $r_{i1} = \cdots = r_{iT} = 1$ and we require that r_{it} is independent of α_i and $u_{i1}, \ldots, u_{iT}$. In these cases, the usual consistency properties of the estimators are not affected if we restrict attention to the available or complete observations only. If selection depends upon the equations' error terms, the OLS, random effects and fixed effects estimators may suffer from selection bias (compare Chapter 7). Subsection 10.8.2 provides additional details on this issue, including some simple tests. In cases with selection bias, alternative estimators have to be used, which are typically computationally unattractive. This is discussed in Subsection 10.8.3. Additional details and discussion on incomplete panels and selection bias can be found in Verbeek and Nijman (1992a, 1996), and Baltagi and Song (2006).

10.8.1 Estimation with Randomly Missing Data

The expressions for the fixed and random effects estimators are easily extended to the unbalanced case. The fixed effects estimator, as before, can be determined as the OLS estimator in the linear model where each individual has its own intercept term. Alternatively, the resulting estimator for β can be obtained directly by applying OLS to the within transformed model, where now all variables are in deviation from the mean *over the available observations*. Individuals that are observed only once provide no information on β and should be discarded in estimation. Defining 'available means' as[24]

$$\bar{y}_i = \frac{\sum_{t=1}^T r_{it} y_{it}}{\sum_{t=1}^T r_{it}}; \quad \bar{x}_i = \frac{\sum_{t=1}^T r_{it} x_{it}}{\sum_{t=1}^T r_{it}},$$

the fixed effects estimator can be concisely written as

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^N \sum_{t=1}^T r_{it}(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^N \sum_{t=1}^T r_{it}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i). \quad (10.89)$$

That is, all sums are simply over the available observations only.

In a similar way, the random effects estimator can be generalized. The random effects estimator for the unbalanced case can be obtained from

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^N \sum_{t=1}^T r_{it}(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' + \sum_{i=1}^N \psi_i T_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \right)^{-1}$$
$$\times \left(\sum_{i=1}^N \sum_{t=1}^T r_{it}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) + \sum_{i=1}^N \psi_i T_i (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) \right), \quad (10.90)$$

where $T_i = \sum_{t=1}^T r_{it}$ denotes the number of periods individual i is observed and

$$\psi_i = \frac{\sigma_u^2}{\sigma_u^2 + T_i \sigma_\alpha^2}.$$

Alternatively, it is obtained by applying OLS to the following transformed model:

$$(y_{it} - \vartheta_i \bar{y}_i) = \beta_0(1 - \vartheta_i) + (x_{it} - \vartheta_i \bar{x}_i)'\beta + v_{it}, \quad (10.91)$$

where $\vartheta_i = 1 - \psi_i^{1/2}$. Note that the transformation applied here is individual specific as depends upon the number of observations for individual i.

Essentially, the more general formulae for the fixed effects and random effects estimators are characterized by the fact that all summations and means are over the available observations only and that T_i replaces T. Completely analogous adjustments apply to the expressions for the covariance matrices of the two estimators given in

[24] We assume that $\sum_{t=1}^T r_{it} \geq 1$, i.e. each individual is observed at least once.

(10.13) and (10.26). Consistent estimators for the unknown variances σ_α^2 and σ_u^2 are given by

$$\hat{\sigma}_u^2 = \frac{1}{\sum_{i=1}^N T_i - N} \sum_{i=1}^N \sum_{t=1}^T r_{it} \left(y_{it} - \bar{y}_i - (x_{it} - \bar{x}_i)'\hat{\beta}_{FE} \right)^2 \quad (10.92)$$

and

$$\hat{\sigma}_\alpha^2 = \frac{1}{N} \sum_{i=1}^N \left[(\bar{y}_i - \hat{\beta}_{0B} - \bar{x}_i'\hat{\beta}_B)^2 - \frac{1}{T_i}\hat{\sigma}_u^2 \right] \quad (10.93)$$

respectively, where $\hat{\beta}_B$ is the between estimator for β, and $\hat{\beta}_{0B}$ is the between estimator for the intercept (both computed as the OLS estimator in (10.21), where the means now reflect 'available means'). Because the efficiency of the estimators for σ_α^2 and σ_u^2 asymptotically has no impact on the efficiency of the random effects estimator, it is possible to use computationally simpler estimators for σ_α^2 and σ_u^2 that are consistent. For example, one could use the standard estimators computed from the residuals obtained from estimating with the balanced subpanel only, and then use (10.90) or (10.91) to compute the random effects estimator.

10.8.2 Selection Bias and Some Simple Tests

In addition to the usual conditions for consistency of the random effects and fixed effects estimator, based on either the balanced subpanel or the unbalanced panel, it was assumed above that the response indicator variable r_{it} was independent of all unobservables in the model. This assumption may be unrealistic. For example, explaining the performance of hedge funds may suffer from the fact that funds with a bad performance are less likely to survive (Baquero, ter Horst and Verbeek, 2005), analysing the effect of an income policy experiment may suffer from biases if people that benefit less from the experiment are more likely to drop out of the panel (Hausman and Wise, 1979) or estimating the impact of the unemployment rate on individual wages may be disturbed by the possibility that people with relatively high wages are more likely to leave the labour market in case of increasing unemployment (Keane, Moffitt and Runkle, 1988).

If r_{it} depends upon α_i or u_{it}, **selection bias** may arise in the standard estimators (see Chapter 7). This means that the distribution of y given x and conditional upon selection (into the sample) is different from the distribution of y given x (which is what we are interested in). For consistency of the fixed effects estimator it is now required that

$$E\{(x_{it} - \bar{x}_i)u_{it} | r_{i1}, \ldots, r_{iT}\} = 0. \quad (10.94)$$

This means that the fixed effects estimator is inconsistent if whether an individual is in the sample or not tells us something about the expected value of the error term that is related with x_{it}. Clearly, if (10.11) holds and r_{it} is independent of α_i and all u_{is} (for given x_{is}), the above condition is satisfied. Note that sample selection may depend upon α_i without affecting consistency of the fixed effects estimator for β. In fact, u_{it}

may even depend upon r_{it} as long as their relationship is time invariant (see Verbeek and Nijman, 1992a, 1996 for additional details).

In addition to (10.94), the conditions for consistency of the random effects estimator are now given by $E\{\bar{x}_i u_{it}|r_{i1},\ldots,r_{iT}\} = 0$ and

$$E\{\bar{x}_i \alpha_i|r_{i1},\ldots,r_{iT}\} = 0. \tag{10.95}$$

This does not allow the expected value of either error component to depend on the selection indicators. If individuals with certain values for their unobserved heterogeneity α_i are less likely to be observed in some wave of the panel, this will typically bias the random effects estimator. Similarly, if individuals with certain shocks u_{it} are more likely to drop out, the random effects estimator is typically inconsistent. Note that, because the fixed effects estimator allows selection to depend upon α_i and upon u_{it} in a time-invariant way, it is more robust against selection bias than the random effects estimator. Another important observation made by Verbeek and Nijman (1992a) is that estimators from the unbalanced panel do not necessarily suffer less from selection bias than those from the balanced subpanel. In general, the selection biases in the estimators from the unbalanced and balanced samples need not be the same, and their relative magnitude is not known a priori.

Verbeek and Nijman (1992a) suggest a number of simple tests for selection bias based upon the above observations. First, as the conditions for consistency state that the error terms should – in one sense or another – not depend upon the selection indicators, one can test this by simply including some function of $r_{i1},\ldots,r_{iT}$ in the model and checking its significance. Clearly, the null hypothesis says that whether an individual was observed in any of the periods 1 to T should not give us any information about his or her unobservables in the model. Obviously, adding r_{it} to the model in (10.88) leads to multicollinearity as $r_{it} = 1$ for all observations in the sample. Instead, one could add functions of $r_{i1},\ldots,r_{iT}$, like $r_{i,t-1}$, $c_i = \prod_{t=1}^{T} r_{it}$ or $T_i = \sum_{t=1}^{T} r_{it}$, indicating whether unit i was observed in the previous period, whether it was observed over all periods and the total number of periods unit i is observed respectively. Note that in the balanced subpanel all variables are identical for all individuals and thus incorporated in the intercept term. Verbeek and Nijman (1992a) suggest that the inclusion of c_i and T_i may provide a reasonable procedure to check for the presence of selection bias. Note that this requires that the model be estimated under the random effects assumption, as the within transformation would wipe out both c_i and T_i. Of course, if the tests do not reject, there is no reason to accept the null hypothesis of no selection bias, because the power of the tests may be low.

Another group of tests is based upon the idea that the four different estimators, random effects and fixed effects, using either the balanced subpanel or unbalanced panel, usually all suffer differently from selection bias. A comparison of these estimators may therefore give an indication for the likelihood of selection bias. Although any pair of estimators can be compared (see Verbeek and Nijman, 1992a, or Baltagi, 2005, Section 11.4), it is known that fixed effects and random effects estimators may be different for other reasons than selection bias (see Subsection 10.2.4). Therefore, it is most natural to compare either the fixed effects or the random effects estimator using the balanced subpanel, with its counterpart using the unbalanced panel. If different samples, selected on the basis of $r_{i1},\ldots,r_{iT}$, lead to significantly different estimators,

it must be the case that the selection process tells us something about the unobservables in the model. That is, it indicates the presence of selection bias. As the estimators using the unbalanced panel are efficient within a particular class of estimators, we can use the result of Hausman again and derive a test statistic based upon the random effects estimator as (compare (10.28))

$$\xi_{RE} = (\hat{\beta}_{RE}^B - \hat{\beta}_{RE}^U)'[\hat{V}\{\hat{\beta}_{RE}^B\} - \hat{V}\{\hat{\beta}_{RE}^U\}]^{-1}(\hat{\beta}_{RE}^B - \hat{\beta}_{RE}^U), \tag{10.96}$$

where the $\hat{V}$s denote estimates of the covariance matrices and the superscripts B and U refer to the balanced and unbalanced sample respectively. Similarly, a test based on the two fixed effects estimators can be derived. Under the null hypothesis, the test statistic follows a Chi-squared distribution with K degrees of freedom. Note that the implicit null hypothesis for the test is that $\text{plim}(\hat{\beta}_{RE}^B - \hat{\beta}_{RE}^U) = 0$. If this is approximately true and the two estimators suffer similarly from selection bias, the test has no power.[25] Again, it is possible to test for a subset of the elements in β.

10.8.3 Estimation with Nonrandomly Missing Data

As in the cross-sectional case (see Section 7.6), selection bias introduces an identification problem. As a result, it is not possible to obtain consistent estimators for the model parameters in the presence of selection bias, unless additional assumptions are imposed. As an illustration, let us assume that the selection indicator r_{it} can be explained by a random effects probit model, that is

$$r_{it}^* = z_{it}'\gamma + \xi_i + \eta_{it}, \tag{10.97}$$

where $r_{it} = 1$ if $r_{it}^* > 0$ and 0 otherwise, and z_{it} is a (well-motivated) vector of exogenous variables that includes x_{it}. The model of interest is given by

$$y_{it} = x_{it}'\beta + \alpha_i + u_{it}. \tag{10.98}$$

Let us assume that the error components in the two equations have a joint normal distribution. This is a generalization of the cross-sectional sample-selection model considered in Subsection 7.6.1. The effect of sample selection in (10.98) is reflected in the expected values of the unobservables, conditional upon the exogenous variables and the selection indicators, that is

$$E\{\alpha_i|z_{i1},\ldots,z_{iT},r_{i1},\ldots,r_{iT}\} \tag{10.99}$$

and

$$E\{u_{it}|z_{i1},\ldots,z_{iT},r_{i1},\ldots,r_{iT}\}. \tag{10.100}$$

It can be shown (Verbeek and Nijman, 1992a) that (10.100) is time invariant if $\text{cov}\{u_{it},\eta_{it}\} = 0$ or if $z_{it}'\gamma$ is time invariant. This is required for consistency of the fixed

[25] The test suggested here is not a real Hausman test because none of the estimators is consistent under the alternative hypothesis. This does not invalidate the test as such but may result in limited power in certain directions.

effects estimator. Further, (10.99) is zero if $\text{cov}\{\alpha_i, \xi_i\} = 0$, while (10.100) is zero if $\text{cov}\{u_{it}, \eta_{it}\} = 0$, so that the random effects estimator is consistent if the unobservables in the primary equation and the selection equation are uncorrelated.

Estimation in the more general case is relatively complicated. Hausman and Wise (1979) consider a case where the panel has two periods and attrition only takes place in the second period. In the more general case, using maximum likelihood to estimate the two equations simultaneously requires numerical integration over two dimensions (to integrate out the two individual effects). Nijman and Verbeek (1992) and Vella and Verbeek (1999a) present alternative estimators based upon the two-step estimation method for the cross-sectional sample-selection model. Essentially, the idea is that the terms in (10.99) and (10.100), apart from a constant, can be determined from the probit model in (10.97), so that estimates of these terms can be included in the primary equation. Wooldridge (1995) presents some alternative estimators based on somewhat different assumptions. Das (2004) extends these approaches to cover flexible functional forms in both (10.97) and (10.98) and unknown distributions for the unobserved components.

Identification of (10.98) with attrition or selection bias using the approaches discussed above depends crucially upon the availability of one or more instruments in (10.97). That is, the variables in z_{it} that are not included in (10.98) should be orthogonal to the unobservables in α_i and (most importantly) u_{it}. In this case, the occurrence of selection bias is driven by the correlations between the unobservables in both equations, a case which is sometimes referred to as 'selection upon unobservables'. An alternative approach to handle nonrandom attrition in panel data requires that z_{it} in (10.97) can be chosen in such a way that the unobservables ξ_i and η_{it} are unrelated to the unobservables in (10.98), while z_{it} may depend upon α_i and u_{it}. This says that a (potentially large) set of observables can be found that are relevant for the selection process such that, conditional upon those variables, selection no longer depends upon the unobservables in (10.98). This case is referred to as 'selection upon observables' and is exploited in Fitzgerald, Gottschalk and Moffitt (1998) to evaluate attrition bias in the Panel Study of Income Dynamics (PSID). In their case, z_{it} contains all available lags of y_{it}. Consistent estimation of (10.98) is achieved by attaching weights to each observation in the panel, where the weights depend upon the selection probability (propensity score). Because the two approaches impose different identification conditions, they cannot be tested against each other. Hirano, Imbens, Ridder and Rubin (2001) show how the availability of refreshment samples (new units randomly sampled from the original population) can be used to distinguish between selection upon unobservables and selection upon observables.

10.9 Pseudo Panels and Repeated Cross-sections

In many countries there is a lack of genuine panel data where specific individuals or firms are followed over time. However, repeated cross-sectional surveys may be available, where a random sample is taken from the population at consecutive points in time. Important examples of this are the Current Population Survey in the USA and the Family Expenditure Survey in the United Kingdom. While many types of model can be estimated on the basis of a series of independent cross-sections in a standard way, several models that seemingly require the availability of panel data can also be identified with repeated cross-sections under appropriate conditions. Most importantly, this concerns models with individual dynamics and models with fixed individual-specific effects.

Obviously, the major limitation of repeated cross-sectional data is that the same individuals are not followed over time, so that individual histories are not available for inclusion in a model, for constructing instruments or for transforming a model to first-differences or in deviations from individual means. All of these are often applied with genuine panel data. On the other hand, repeated cross-sections suffer much less from typical panel data problems like attrition and nonresponse, and are very often substantially larger, both in number of individuals or households and in the time period that they span.

10.9.1 The Fixed Effects Model

Consider the linear model with individual effects given by

$$y_{it} = x_{it}'\beta + \alpha_i + u_{it}, \quad t = 1, \ldots, T. \tag{10.101}$$

Unlike the previous sections, the available data set is a series of independent cross-sections, such that observations on N different individuals are available in each period.[26] For simplicity, we shall assume that $E\{x_{it}u_{it}\} = 0$ for each t. If the individual effects α_i are uncorrelated with the explanatory variables in x_{it}, the model in (10.101) can easily be estimated consistently from repeated cross-sections by pooling all observations and performing ordinary least squares treating $\alpha_i + u_{it}$ as a composite error term and including an overall intercept term. However, in many applications the individual effects are likely to be correlated with some or all of the explanatory variables, and OLS is inconsistent. When genuine panel data are available, this can be solved using the within or first-difference transformation to eliminate α_i. Obviously, when repeated observations on the same individuals are not available, such an approach cannot be used.

Deaton (1985) suggests the use of cohorts to obtain consistent estimators for β in (10.101) when repeated cross-sections are available, even if α_i is correlated with one or more of the explanatory variables. Let us define C cohorts, which are groups of individuals sharing some common characteristics. These groups are defined such that each individual is a member of exactly one cohort, which is the same for all periods. For example, a particular cohort can consist of all males born in the period 1950–1954. It is important to realize that the variables on which cohorts are defined should be observed for all individuals in the sample. This rules out time-varying variables (e.g. earnings), because these variables are observed at different points in time for the individuals in the sample. The seminal study of Browning, Deaton and Irish (1985) employs cohorts of households defined on the basis of 5 year age bands subdivided as to whether the head of the household is a manual or nonmanual worker.

[26] Because different individuals are observed in each period, this implies that i does not run from 1 to N for each t.

If we aggregate all observations to cohort level, the resulting model can be written as

$$\bar{y}_{ct} = \bar{x}'_{ct}\beta + \bar{\alpha}_{ct} + \bar{u}_{ct}, \quad c = 1, \ldots, C; \quad t = 1, \ldots, T, \tag{10.102}$$

where $\bar{y}_{ct}$ is the average value of all observed y_{it}s in cohort c in period t, and similarly for the other variables in the model. The resulting data set is a **pseudo panel** or synthetic panel with repeated observations over T periods and C cohorts. The main problem with estimating β from (10.102) is that $\bar{\alpha}_{ct}$ depends on t, is unobserved and is likely to be correlated with $\bar{x}_{ct}$ (if α_i is correlated with x_{it}). Therefore, treating $\bar{\alpha}_{ct}$ as part of the random error term is likely to lead to inconsistent estimators. Alternatively, one can treat $\bar{\alpha}_{ct}$ as fixed unknown parameters assuming that variation over time can be ignored ($\bar{\alpha}_{ct} = \alpha_c$). If cohort averages are based on a large number of individual observations, this assumption seems reasonable and a natural estimator for β is the within estimator on the pseudo panel, given by

$$\hat{\beta}_W = \left(\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)'\right)^{-1} \sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)(\bar{y}_{ct} - \bar{y}_c), \tag{10.103}$$

where $\bar{x}_c = T^{-1}\sum_{t=1}^{T}\bar{x}_{ct}$ is the time average of the observed cohort means for cohort c. The properties of this estimator depend, among other things, upon the type of asymptotics that one is willing to employ. In addition to the two dimensions in genuine panel data (N and T), there are two additional dimensions: the number of cohorts C and the number of observations per cohort n_c. A convenient choice is to let $N \to \infty$, with C fixed, so that $n_c \to \infty$. Then the fixed effects estimator based on the pseudo panel, $\hat{\beta}_W$, is consistent for β, provided that

$$\operatorname*{plim}_{n_c \to \infty} \frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)' \tag{10.104}$$

is finite and invertible, and that

$$\operatorname*{plim}_{n_c \to \infty} \frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)\bar{\alpha}_{ct} = 0. \tag{10.105}$$

While the first of these two conditions is similar to a standard regularity condition (compare assumption (A6) in Section 2.6), in this context it is somewhat less innocent. It states that the cohort averages exhibit genuine time variation, even with very large cohorts. Whether or not this condition is satisfied depends upon the way the cohorts are constructed, a point to which we shall return below.

Because $\bar{\alpha}_{ct} \to \alpha_c$, for some α_c, if the number of observations per cohort tends to infinity, (10.105) will be satisfied automatically. Consequently, letting $n_c \to \infty$ is a convenient choice to arrive at a consistent estimator for β; see Moffitt (1993) and Ridder and Moffitt (2007). However, as argued by Verbeek and Nijman (1992b) and Devereux (2007), even if cohort sizes are large, the small-sample bias in the within estimator on the pseudo panel may still be substantial. Deaton (1985) considers alternative errors-in-variables estimators for β that do not depend upon $n_c \to \infty$ but instead impose that $N \to \infty$ and $C \to \infty$, with n_c fixed.

10.9.2 An Instrumental Variables Interpretation

To appreciate the role of the way in which the cohorts are constructed, it is useful to reformulate the above estimator as an instrumental variables estimator based on a simple extension of equation (10.101). The idea advocated by Moffitt (1993) is that grouping can be viewed as an instrumental variables procedure. First, decompose each individual effect α_i into a cohort effect α_c and individual i's deviation from this effect. Letting $z_{ci} = 1$ $(c = 1, \ldots, C)$ if individual i is a member of cohort c and 0 otherwise, we can write

$$\alpha_i = \sum_{c=1}^{C}\alpha_c z_{ci} + v_i, \tag{10.106}$$

which can be interpreted as an orthogonal projection. Defining $\alpha = (\alpha_1, \ldots, \alpha_C)'$ and $z_i = (z_{1i}, \ldots, z_{Ci})'$ and substituting (10.106) into (10.101), we obtain

$$y_{it} = x'_{it}\beta + z'_i\alpha + v_i + u_{it}. \tag{10.107}$$

If α_i and x_{it} are correlated, we may also expect that v_i and x_{it} are correlated. Consequently, estimating (10.107) by ordinary least squares would not result in consistent estimators. Now, suppose that instruments for x_{it} can be found that are uncorrelated with $v_i + u_{it}$. In this case, an instrumental variables estimator would typically produce a consistent estimator for β and α_c. A natural choice is to choose the cohort dummies in z_i, interacted with time, as instruments, in which case we derive linear predictors from the K reduced forms:

$$x_{k,it} = z'_i\delta_{kt} + w_{k,it}, \quad k = 1, \ldots, K, t = 1, \ldots, T, \tag{10.108}$$

where δ_{kt} is a vector of unknown parameters. The linear predictor for x_{it} by construction equals $\bar{x}_{ct}$, the vector of averages within cohort c in period t. The resulting instrumental variables estimator for β is then given by

$$\hat{\beta}_{IV1} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)x'_{it}\right)^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)y_{it}, \tag{10.109}$$

which is identical to the standard within estimator based on the pseudo panel of cohort averages, given in (10.103).

The instrumental variables interpretation is useful because it illustrates that alternative estimators may be constructed using other sets of instruments. For example, z_i may include (smooth) functions of year of birth, rather than a set of dummy variables. Further, the instrument set in (10.108) can be extended to include additional variables. Most importantly, however, the instrumental variables approach stresses that grouping data into cohorts requires grouping variables that should satisfy the typical requirements for instrument exogeneity and relevance.

In practice, cohorts should be defined on the basis of variables that do not vary over time and that are observed for all individuals in the sample. This is a serious restriction. Possible choices include variables like age (date of birth), gender, race or region.[27] Identification of the parameters in the model requires that the reduced forms in (10.108) generate sufficient variation over time. This requirement puts a heavy burden on the cohort identifying variables. In particular, it requires that groups are defined whose explanatory variables all have changed differentially over time. Suppose, as an extreme example, that cohorts are defined on the basis of a variable that is independent of the variables in the model. That is, cohorts are constructed by randomly grouping individuals. In this case, the true population cohort means x_{ct} would be identical for each cohort c (and equal the overall population mean). This leaves only the time variation in x_{ct} to identify the parameters of interest.

10.9.3 Dynamic Models

An important situation where the availability of panel data seems essential to identify and estimate the model of interest is the case where a lagged dependent variable enters the model. Let us consider a simple extension of (10.101) given by

$$y_{it} = \gamma y_{i,t-1} + x_{it}'\beta + \alpha_i + u_{it}, \quad t = 1, \ldots, T, \tag{10.110}$$

where the K-dimensional vector x_{it} may include time-invariant and time-varying variables. When genuine panel data are available, the parameters γ and β can be estimated consistently (for fixed T and $N \to \infty$) using the instrumental variables estimators and GMM estimators discussed in Section 10.4. These estimators are based on first-differencing (10.110) and then using lagged values of $y_{i,t-1}$ as instruments.

In the present context, $y_{i,t-1}$ refers to the value of y at $t-1$ for an individual who is only observed in cross-section t. Thus, an observation for $y_{i,t-1}$ is unavailable. Therefore, the first step is to construct an estimate by using information on the y values of other individuals observed at $t-1$. A convenient approach is to use the average value of $y_{i,t-1}$ from individuals in the same cohort, $\bar{y}_{c,t-1}$, say. Inserting these predicted values into the original model, we obtain

$$y_{it} = \gamma \bar{y}_{c,t-1} + x_{it}'\beta + \xi_{i,t}, \quad t = 1, \ldots, T, \tag{10.111}$$

where

$$\xi_{it} = \alpha_i + u_{it} + \gamma(y_{i,t-1} - \bar{y}_{c,t-1}). \tag{10.112}$$

The unobserved prediction error $y_{i,t-1} - \bar{y}_{c,t-1}$ is part of the error term and is also likely to be correlated with x_{it}. As a result, OLS estimation of (10.111) is typically inconsistent (see Verbeek and Vella, 2005, for more discussion and exceptions). To overcome this problem, one can use an instrumental variables approach. Note that now we need instruments for x_{it} even though these variables are exogenous in the original model. As before, a natural choice is to use the cohort dummies, interacted with time, as instruments for x_{it}. These instruments are uncorrelated with $y_{i,t-1} - \bar{y}_{c,t-1}$ by construction.

When the instruments z_i are a set of cohort dummies, estimation of (10.111) by instrumental variables is identical to applying OLS to the original model where all

[27] Note that residential location may be endogenous in certain applications.

variables are replaced by their (time-specific) cohort sample averages. We can write this as

$$\bar{y}_{ct} = \gamma \bar{y}_{c,t-1} + \bar{x}_{ct}'\beta + \bar{\xi}_{ct}, \quad c = 1, \ldots, C, \quad t = 1, \ldots, T, \tag{10.113}$$

where all variables denote period-by-period averages within each cohort. For this approach to be appropriate, we need $\bar{y}_{c,t-1}$ and $\bar{x}_{ct}$ not to be collinear, which requires the instruments to capture variation in $y_{i,t-1}$ independently of the variation in x_{it}. It is possible to include cohort fixed effects in essentially the same way as in the static linear model by including the cohort dummies in the equation of interest, with time-invariant coefficients. This imposes (10.106) and results in

$$\bar{y}_{ct} = \gamma \bar{y}_{c,t-1} + \bar{x}_{ct}'\beta + \alpha_c + \bar{u}_{c,t}, \tag{10.114}$$

where α_c denotes a cohort-specific fixed effect. Applying OLS to (10.114) corresponds to the standard within estimator for γ and β based upon treating the cohort-level data as a panel, which is consistent under the given assumptions (and some regularity conditions) when $n_c \to \infty$ and C is fixed. The usual problem with estimating dynamic panel data models with short T (see Section 10.4), does not arise because the error term, which is a within cohort average of individual error terms that are uncorrelated with z_i, is asymptotically zero.[28] However, it remains to be seen whether suitable instruments can be found that satisfy the above conditions, because the rank condition for identification requires that the time-invariant instruments have time-varying relationships with the exogenous variables and the lagged dependent variable, while they should not have any time-varying relationship with the equation's error term. While this seems unlikely, it is not impossible. When z_i is uncorrelated with u_{it}, it is typically sufficient that the means of the exogenous variables, conditional upon z_i, are time varying; see Verbeek and Vella (2005) for more details.

McKenzie (2004) considers the linear dynamic model with cohort-specific coefficients in equation (10.110). While this extension will typically only make sense if there is a fairly small number of well-defined cohorts, it arises naturally from the existing literature on dynamic heterogeneous panels. For example, Robertson and Symons (1992) and Pesaran and Smith (1995) stress the importance of parameter heterogeneity in dynamic panel data models and analyse the potentially severe biases that may arise from handling it in an inappropriate manner. In many practical applications, investigating whether there are systematic differences between, for example, age cohorts is an interesting question. Obviously, relaxing specification (10.110) by having cohort-specific coefficients puts an additional burden upon the identifying conditions. Verbeek (2008) provides additional discussion and references on pseudo panel data.

Exercises

Exercise 10.1 (Linear Model)

Consider the following simple panel data model

$$y_{it} = x_{it}\beta + \alpha_i^* + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{10.115}$$

[28] Recall that, asymptotically, the number of cohorts is fixed and the number of individuals goes to infinity.

where β is one-dimensional, and where it is assumed that

$$\alpha_i^* = \bar{x}_i\lambda + \alpha_i, \quad \text{with} \quad \alpha_i \sim NID(0, \sigma_\alpha^2), \quad u_{it} \sim NID(0, \sigma_u^2).$$

The two error components α_i and u_{it} are mutually independent and independent of all x_{it}s.

The parameter β in (10.101) can be estimated by the fixed effects (or within) estimator given by

$$\hat{\beta}_{FE} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)^2}.$$

As an alternative, the correlation between the error term $\alpha_i^* + u_{it}$ and x_{it} can be handled by an instrumental variables approach.

a. Give an expression for the IV estimator $\hat{\beta}_{IV}$ for β in (10.101) using $x_{it} - \bar{x}_i$ as an instrument for x_{it}. Show that $\hat{\beta}_{IV}$ and $\hat{\beta}_{FE}$ are identical.

Another way to eliminate the individual effects α_i^* from the model is to take first-differences. This results in

$$y_{it} - y_{i,t-1} = (x_{it} - x_{i,t-1})\beta + (u_{it} - u_{i,t-1}), \quad i = 1, \ldots, N, \quad t = 2, \ldots, T. \tag{10.116}$$

b. Denote the OLS estimator based on (10.116) by $\hat{\beta}_{FD}$. Show that $\hat{\beta}_{FD}$ is identical to $\hat{\beta}_{IV}$ and $\hat{\beta}_{FE}$ if $T = 2$. This identity no longer holds for $T > 2$. Which of the two estimators would you prefer in that case? Explain. (Note: for additional discussion, see Verbeek, 1995.)

c. Consider the between estimator $\hat{\beta}_B$ for β in (10.115). Give an expression for $\hat{\beta}_B$ and show that it is unbiased for $\beta + \lambda$.

d. Finally, suppose we substitute the expression for α_i^* into (10.115), giving

$$y_{it} = x_{it}\beta + \bar{x}_i\lambda + \alpha_i + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T. \tag{10.117}$$

The vector $(\beta, \lambda)'$ can be estimated by GLS (random effects) based on (10.117). It can be shown that the implied estimator for β is identical to $\hat{\beta}_{FE}$. Does this imply that there is no real distinction between the fixed effects and random effects approaches? (Note: for additional discussion, see Hsiao, 2003, Section 3.4.2a.)

Exercise 10.2 (Hausman–Taylor Model)

Consider the following linear panel data model:

$$y_{it} = x'_{1,it}\beta_1 + x'_{2,it}\beta_2 + w'_{1,i}\gamma_1 + w'_{2,i}\gamma_2 + \alpha_i + u_{it}, \tag{10.118}$$

where $w_{k,i}$ are time-invariant and $x_{k,it}$ are time-varying explanatory variables. The variables with index 1 ($x_{1,it}$ and $w_{1,i}$) are strictly exogenous in the sense that $E\{x_{1,it}\alpha_i\} = 0$, $E\{x_{1,is}u_{it}\} = 0$ for all s, t, $E\{w_{1,i}\alpha_i\} = 0$ and $E\{w_{1,i}u_{it}\} = 0$. It is also assumed that

$E\{w_{2,i}u_{it}\} = 0$ and that the usual regularity conditions (for consistency and asymptotic normality) are met.

a. Under which additional assumptions would OLS applied to (10.118) provide a consistent estimator for $\beta = (\beta_1, \beta_2)'$ and $\gamma = (\gamma_1, \gamma_2)'$?

b. Consider the fixed effects (within) estimator. Under which additional assumption(s) would it provide a consistent estimator for β?

c. Consider the OLS estimator for β based upon a regression in first-differences. Under which additional assumption(s) will this provide a consistent estimator for β?

d. Discuss one or more alternative consistent estimators for β and γ if it can be assumed that $E\{x_{2,is}u_{it}\} = 0$ (for all s, t), and $E\{w_{2,i}u_{it}\} = 0$. What are the restrictions, in this case, on the number of variables in each of the categories?

e. Discuss estimation of β if $x_{2,it}$ *equals* $y_{i,t-1}$.

f. Discuss estimation of β if $x_{2,it}$ *includes* $y_{i,t-1}$.

g. Would it be possible to estimate both β and γ consistently if $x_{2,it}$ includes $y_{i,t-1}$? If so, how? If not, why not? (Make additional assumptions, if necessary.)

Exercise 10.3 (Linear Model – Empirical)

This exercise makes use of data for young females from the National Longitudinal Survey (Youth Sample) for the period 1980–1987, available from the book's website. These data are also used in Vella and Verbeek (1999a). We focus on the subsample of 12 039 observations reporting positive hours of work in a given period.

a. Produce summary statistics of the data set and produce a histogram of T_i. How many individuals do you have in the panel? How many of them are continuously working over the entire period 1980–1987?

b. Estimate a simple wage equation using pooled OLS, with clustered (panel-robust) standard errors. Explain a person's log wage from marital status, black, hispanic, schooling, experience and experience-squared, rural and union membership. Estimate another specification that includes time dummies. Compare the results. Test whether the time dummies are jointly significant. Why does the inclusion of time dummies make sense economically?

c. Use the fixed effects and random effects estimators to estimate the same equation. Interpret and compare the results. (You may also want to compare the results with those for males reported in Table 10.2.)

d. Perform a Hausman test and interpret the result. What exactly is the null hypothesis that you test?

e. On the basis of the random effects results, interpret the estimates for σ_u^2 and σ_α^2 and use them to estimate the transformation factor ϑ in (10.23). How important is the individual effect in this equation?

f. Re-estimate the wage equation, using the random effects estimator, including age and age-squared rather than experience and experience-squared. Compare the results. What happened to the coefficient on schooling? Why?

g. Let us focus on the random effects model including experience and experience-squared. Re-estimate this model including T_i and interpret the results. Evaluate the t-test on the included variable. What does it test? Does the result surprise you? Why doesn't this test work with the fixed effects model? Repeat the estimation but include a dummy for $T_i = 8$. Interpret.

h. Re-estimate the base model (with experience and experience-squared) from **c** using the random effects estimator, using the unbalanced panel and the balanced subpanel (characterized by $T_i = 8$). Compare the results. Does it appear that the loss in efficiency is substantial? What about the coefficient estimates?

i. Perform a Hausman test on the difference between the two estimators in **h** and interpret the results.

j. Repeat the previous test using the fixed effects estimator. Interpret and compare with **i**. If you experience problems calculating the Hausman test statistic, try using panel-robust covariance matrices.

Exercise 10.4 (Dynamic and Binary Choice Models)

Consider the following dynamic wage equation

$$w_{it} = x_{it}'\beta + \gamma w_{i,t-1} + \alpha_i + u_{it}, \tag{10.119}$$

where w_{it} denotes an individual's log hourly wage rate and x_{it} is a vector of personal and job characteristics (age, schooling, gender, industry, etc.).

a. Explain in words why OLS applied to (10.119) is inconsistent.

b. Also explain why the fixed effects estimator applied to (10.119) is inconsistent for $N \to \infty$ and fixed T, but consistent for $N \to \infty$ and $T \to \infty$. (Assume that u_{it} is i.i.d.)

c. Explain why the results from **a** and **b** also imply that the random effects (GLS) estimator in (10.119) is inconsistent for fixed T.

d. Describe a simple consistent (for $N \to \infty$) estimator for β, γ, assuming that α_i and u_{it} are i.i.d. and independent of all x_{it}s.

e. Describe a more efficient estimator for β, γ under the same assumptions.

In addition to the wage equation, assume there is a binary choice model explaining whether an individual is working or not. Let $r_{it} = 1$ if individual i was working in period t and zero otherwise. Then the model can be described as

$$\begin{aligned} r_{it}^* &= z_{it}'\delta + \xi_i + \eta_{it} \\ r_{it} &= 1 \quad \text{if } r_{it}^* > 0 \\ &= 0 \quad \text{otherwise,} \end{aligned} \tag{10.120}$$

where z_{it} is a vector of personal characteristics. Assume that $\xi_i \sim NID(0, \sigma_\xi^2)$ and $\eta_{it} \sim NID(0, 1 - \sigma_\xi^2)$, mutually independent and independent of all z_{it}s. The model in (10.120) can be estimated by maximum likelihood.

f. Give an expression for the probability that $r_{it} = 1$ given z_{it} and ξ_i.

g. Use the expression from **f** to obtain a computationally tractable expression for the likelihood contribution of individual i.

h. Explain why it is not possible to treat the ξ_is as fixed unknown parameters and estimate δ consistently (for fixed T) from this fixed effects probit.

From now on, assume that the appropriate wage equation is static and given by (10.119) with $\gamma = 0$.

i. What are the consequences for the random effects estimator in (10.119) if η_{it} and u_{it} are correlated? Why?

j. What are the consequences for the fixed effects estimator in (10.119) if ξ_i and α_i are correlated (while η_{it} and u_{it} are not)? Why?

Exercise 10.5 (Binary Choice Models – Empirical)

This exercise makes use of data for young females from the National Longitudinal Survey (Youth Sample) for 1980–1987, also used in Exercise 10.3. Our goal is to model union status of working females.

a. Produce summary statistics for union status. How many observations relate to union members? How many females are union members for all periods they are in the panel? How many females are never union members?

b. Estimate a pooled probit model (ignoring the panel nature of the data) explaining union status from age, schooling, hispanic, black, public sector, marital status and a dummy for living in the North East. Interpret the results. Is this estimator consistent? What about its standard errors?

c. Re-estimate the pooled probit using panel-robust standard errors. Compare the results with **b** and interpret.

d. Estimate a pooled logit model explaining union status from the same explanatory variables, also with panel-robust standard errors. Compare the estimated coefficients and their significance with those obtained in **c**. Why are the logit coefficients uniformly bigger than the probit ones?

e. Estimate a random effects probit model based on the previous specification. Can you explain why it is taking so much time to determine the maximum likelihood estimates for this model? Interpret the estimation results. Also report which normalization constraint is imposed upon σ_α^2 and σ_u^2. Use this to compare the coefficient estimates from the random effects probit model with those from the pooled probit model.

f. Perform a likelihood ratio test on the restriction that $\sigma_\alpha^2 = 0$. Interpret.

g. Extend the previous model with a lagged dependent variable (lagged union status). Compare the estimation results with those obtained under **e**. Also compare the estimated value of σ_α^2. Explain. Under what conditions is it appropriate to include a lagged dependent variable in a random effects binary choice model? Are you concerned with the fact that the estimated autoregressive coefficient is bigger than one?

h. Estimate a static fixed effects logit model. Interpret the results. How many individuals are used to estimate this model?

A Guide to Modern Econometrics

3rd edition

Marno Verbeek
RSM Erasmus University, Rotterdam

John Wiley & Sons, Ltd