Pseudo panels and repeated cross-sections

Marno Verbeek*

November 12, 2007

Abstract

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. In this paper we discuss the identification and estimation of panel data models from repeated cross sections. In particular, attention will be paid to linear models with fixed individual effects, to models containing lagged dependent variables and to discrete choice models.

*RSM Erasmus University and Netspar, Dept. of Financial Management, P.O. Box 1738, 3000 DR, The Netherlands, mverbeek@rsm.nl. This chapter is prepared for: L. Mátyás and P. Sevestre, eds., (2008), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, to be published by Springer in the Spring of 2008 (ISBN 978-3-540-75889-1).

1 Introduction

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 U.S.A., and the Family Expenditure Survey in the United Kingdom. While many types of models 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.

In a seminal paper, Deaton (1985) suggests the use of cohorts to estimate a fixed effects model from repeated cross-sections. In his approach, individuals sharing some common characteristics (most notably year of birth) are grouped into cohorts, after which the averages within these cohorts are treated as observations in a pseudo panel. Moffitt (1993) and Collado (1997), in different ways, extend the approach of Deaton to nonlinear and dynamic models. Alternative estimators for the model with individual dynamics, including the one proposed by Girma (2000), are evaluated in Verbeek and Vella (2005). Alternative types of asymptotics are discussed in McKenzie (2004). In this chapter we shall discuss the identification and estimation of panel data models from repeated cross sections. In particular, attention will be paid to linear models with fixed individual effects, to models contained lagged dependent variables and to discrete choice models.

Models containing individual effects that are correlated with the explanatory variables ("fixed effects models") often arise naturally from economic theory, for example in life cycle models where the individual effects represent marginal utility of wealth (see, for example, Heckman and MaCurdy, 1980 or Browning, Deaton and Irish, 1985). Individual dynamics also often follow from economic theory, reflecting adjustment costs, habit persistence, or in-

tertemporal optimization. Consequently, from an economic point of view it is important to be able to estimate dynamic models and models with fixed individual effects, even in the absence of genuine panel data. While it is possible to estimate such models using repeated cross-sections, we shall see below that such approaches typically require strong identification conditions, which are often hard to test.

Estimation techniques based on grouping individual data into cohorts are identical to instrumental variables approaches where the group indicators are used as instruments. Consequently, the grouping variables should satisfy the appropriate conditions for an instrumental variables estimator to be consistent (including a rank condition). This not only requires that the instruments are valid (in the sense of being uncorrelated to the unobservables in the equation of interest), but also relevant, i.e. appropriately correlated to the explanatory variables in the model. Loosely speaking, the latter requirement means that cohorts are defined as groups whose explanatory variables change differentially over time. Even if the instruments are theoretically valid and relevant, their large number and the fact that they may be only weakly correlated with the explanatory variables they are supposed to instrument may imply that the resulting estimators perform poorly because of the "weak instruments" problem (see Bound, Jaeger and Baker, 1995, or Staiger and Stock, 1997).

The structure of this chapter is as follows. In Section 2 we present the basic linear model. Section 3 pays attention to linear dynamic models, while Section 4 briefly discusses the estimation of binary choice models. Section 5 concludes. A related survey can be found in Ridder and Moffitt (2007).

2 Estimation of a linear fixed effects model

We start with analyzing a simple linear model with individual effects given by

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

where x_{it} denotes a K-dimensional vector of explanatory variables, and β is the parameter vector of interest. The index i refers to individuals and throughout this chapter we shall assume that the available data set is a series of independent cross-sections, such that observations on N individuals are available in each period.[1] For simplicity, we shall assume that $E\{x_{it}u_{it}\} = 0$ for each t.

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

If the individual effects α_i are uncorrelated with the explanatory variables in x_{it}, the model in (1) can easily be estimated consistently from repeated cross-sections by pooling all observations and performing ordinary least squares treating $\alpha_i + u_{it}$ as composite error term. This exploits the K moment conditions in

$$E\{(y_{it} - x'_{it}\beta)x_{it}\} = 0. \tag{2}$$

However, in many applications the individual effects are likely to be correlated with some or all of the explanatory variables, so that at least some of the moment conditions in (2) are not valid. When genuine panel data are available, this can be solved using a fixed effects approach which treats α_i as fixed unknown parameters. In other words, each individual has its own intercept term. For estimating β, this is equivalent to using the within-transformed explanatory variables $x_{it} - \bar{x}_i$ as instruments for x_{it} in (1), where $\bar{x}_i = T^{-1} \sum_{t=1}^{T} x_{it}$. 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 (1) 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 may 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 five-year age bands subdivided as to whether the head-of-the-household is a manual or non-manual worker. Blundell, Duncan and Meghir (1998) employ year-of-birth intervals of 10 years, interacted with two education groups, Banks, Blundell and Preston (1994) use five-year age bands, while Propper, Rees and Green (2001) use 7 date of birth groups and 10 regions to construct cohorts.[2]

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

$$\bar{y}_{ct} = \bar{x}'_{ct}\beta + \bar{\alpha}_{ct} + \bar{u}_{ct}, \;\; c = 1, ..., C; \;\; t = 1, ..., T, \tag{3}$$

[2]Some authors employ the term "cohorts" to specifically reflect year-of-birth groups. We use "cohorts" in a broader sense, as groups of individuals (households, firms) sharing some common characteristics (most often including year-of-birth).

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 (3) 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

$$\hat{\beta}_W = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)' \right)^{-1} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{y}_{ct} - \bar{y}_c), \quad (4)$$

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. Deaton (1995) considers the asymptotic properties of this estimator when the number of cohorts C tends to infinity. This requires that the number of individuals N tends to infinity with (more or less) constant cohort sizes. Moffitt (1993), on the other hand, assumes that C is constant while the number of individuals tends to infinity. In this approach, cohort sizes tend to infinity, asymptotically.

The estimators proposed by Moffitt (1993) are based on the idea that grouping can be viewed as an instrumental variables procedure. To illustrate this, we shall reformulate the above estimator as an instrumental variables estimator based on a simple extension of equation (1). First, decompose each individual effect α_i into a cohort effect α_c and individual i's deviation from this effect. Letting $z_{ci} = 1$ $(c = 1, ..., C)$ if individual i is a member of cohort c and 0 otherwise, we can write

$$\alpha_i = \sum_{c=1}^{C} \alpha_c z_{ci} + v_i, \quad (5)$$

which can be interpreted as an orthogonal projection. Defining $\alpha = (\alpha_1, ..., \alpha_C)'$ and $z_i = (z_{1i}, ..., z_{Ci})'$ and substituting (5) into (1), we obtain

$$y_{it} = x_{it}'\beta + z_i'\alpha + v_i + u_{it}. \quad (6)$$

If α_i and x_{it} are correlated, we may also expect that v_i and x_{it} are correlated. Consequently, estimating (6) 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 reduced forms

$$x_{k,it} = z_i'\delta_{kt} + w_{k,it}, \;\; k = 1, ..., K, \;\; t = 1, ..., T, \tag{7}$$

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

$$\hat{\beta}_{IV1} = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c) x_{it}' \right)^{-1} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c) y_{it}, \tag{8}$$

which is identical to the standard within estimator based on the pseudo panel of cohort averages, given in (4).

The instrumental variables interpretation is useful because it illustrates that alternative estimators may be constructed using other sets of instruments. For example, if cohorts are constructed on the basis of age (year of birth), a more parsimonious function of age can be employed in (5) rather than a full set of age dummies. For example, z_i may include functions of year of birth, rather than a set of dummy variables. As argued by Moffitt (1993), it is likely that y_{it} will vary smoothly with cohort effects and, hence, those effects will be representable by fewer parameters than a full set of cohort dummies. Further, the instrument set in (7) 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 validity and relevance. Basically, the approach of Deaton (1985) assumes that the cohort dummies, interacted with time dummies, provide valid instruments for all explanatory variables in the model (including the full set of cohort dummies). This requires that the instruments are uncorrelated with the equation's error term, and imposes a rank condition stating that the instruments are "sufficiently" correlated with each of the explanatory variables.

As mentioned above, the asymptotic behavior of pseudo panel data estimators can be derived using alternative asymptotic sequences. 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. We consider the following possibilities, which are typical for most studies:

1. $N \to \infty$, with C fixed, so that $n_c \to \infty$;
2. $N \to \infty$ and $C \to \infty$, with n_c fixed.
3. $T \to \infty$, with N, C fixed (so that n_c is also fixed);

McKenzie (2004) also considers asymptotic sequences where $T \to \infty$ and $n_c \to \infty$. Note that asymptotic theory is not meant as a guideline for how our estimators will behave when we get more data. Rather, we appeal to asymptotic theory when some dimension of the sample we already have is large enough for this to be appropriate. Whether or not asymptotic theory provides a reasonable approximation of the finite sample properties of pseudo panel data estimators is an empirical question, and many papers present Monte Carlo studies to obtain some insight into this issue.

The following list provides an overview of the sample sizes used in several important empirical papers.

	T	C	$\bar{n}_c$
Browning, Deaton and Irish (1985)	7	16	190
Banks, Blundell and Preston (1994)	20	11	354
Blundell, Browning and Meghir (1994)	17	9	520
Alessie, Devereux and Weber (1997)	14	5	>1000
Blundell, Duncan and Meghir (1998)	25	8	142
Propper, Rees and Green (2001)	19	70	80

For most applications either type 1 or type 2 asymptotics provides the most reasonable choice, and in many cases type 1 asymptotics is (implicitly or explicitly) employed. In the theoretical literature, Moffitt (1993) and Verbeek and Vella (2005) employ type 1 asymptotics, while Deaton (1985), Verbeek and Nijman (1993) and Collado (1997) employ type 2 (with or without $T \to \infty$). Under type 1 asymptotics, the fixed effects estimator based on the pseudo panel, $\hat{\beta}_W$, is consistent for β, provided that

$$\operatorname*{plim}_{n_c \to \infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)' \tag{9}$$

is finite and invertible, and that

$$\operatorname*{plim}_{n_c \to \infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)\bar{\alpha}_{ct} = 0. \tag{10}$$

While the first of these two conditions is similar to a standard regularity condition, 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) will be satisfied automatically. Consequently, letting $n_c \to \infty$ and using type 1 asymptotics 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 (1992) 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) proposes an alternative estimator for β that does not reply upon having a large number of observations per cohort, using type 2 asymptotics.[3] A convenient starting point for this estimator is the cohort population version of (3), given by

$$y_{ct} = x'_{ct}\beta + \alpha_c + u_{ct}, \;\; c = 1, ..., C; \;\; t = 1, ..., T, \tag{11}$$

where the variables denote unobservable population cohort means, and where α_c is the cohort fixed effect, which is constant because population cohorts contain the same individuals in each period. Now, $\bar{x}_{ct}$ and $\bar{y}_{ct}$ can be considered as error-ridden measurements of x_{ct} and y_{ct}. In particular, it is assumed that the measurement errors are distributed with zero mean, independent of the true values, i.e.

$$\begin{pmatrix} \bar{y}_{ct} - y_{ct} \\ \bar{x}_{ct} - x_{ct} \end{pmatrix} \sim IID\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} \sigma_{00} & \sigma' \\ \sigma & \Sigma \end{pmatrix} \right), \tag{12}$$

where the population cohort means are treated as fixed unknown constants. Although Σ, σ and σ_{00} are unknown, they can easily be estimated consistently (for N or T tending to infinity), using the individual data. Once estimates for Σ and σ are available, it is easy to adjust the moment matrices in the within estimator to eliminate the variance due to measurement error (cf. Fuller,

[3] As argued by McKenzie (2004), in many applications cohorts are defined by age groups and hence a fixed number of cohorts is most likely to be of interest, which is inconsistent with type 2 asymptotics. If $C \to \infty$ with $N \to \infty$, one needs to think of what this means for the distribution of population cohort means as well as the distribution of individual observations around these means. For example, it would be hard to argue that the covariance matrix on the right-hand side of (12) below is independent of how many cohorts are distinguished. See Verbeek and Nijman (1992) for more discussion and a Monte Carlo experiment that takes this issue into account.

1987). This leads to the following errors-in-variables estimator

$$\hat{\beta}_D = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{x}_{ct} - \bar{x}_c)' - \tau \hat{\Sigma} \right)^{-1} \times \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{y}_{ct} - \bar{y}_c)' - \tau \hat{\sigma} \right), \tag{13}$$

where $\hat{\Sigma}$ and $\hat{\sigma}$ are estimates of Σ and σ, respectively, and where $\tau = (T - 1)/T$. As discussed in Verbeek and Nijman (1993), the original estimator presented by Deaton (1995) is characterized by $\tau = 1$. However, eliminating the incidental parameters (in α_c) first by within transforming the data, and working out the appropriate moments, suggests $\tau = (T - 1)/T$, which leads to better small sample properties.

Under type 1 asymptotics, the number of observations per cohort tends to infinity and both Σ and σ tend to zero, as well as their estimators. In this case $\hat{\beta}_D$ is asymptotically equivalent to $\hat{\beta}_W$. Accordingly, most empirical studies ignore the errors-in-variables problem and use standard estimators, like $\hat{\beta}_W$, see, for example, Browning, Deaton and Irish (1985), with an average cohort size of 190, or Blundell, Browning and Meghir (1994), with cohort sizes around 500. Unfortunately, there is no general rule to judge whether n_c is large enough to use asymptotics based on $n_c \to \infty$. Verbeek and Nijman (1992) analyze the bias in $\hat{\beta}_W$ for finite values of n_c. Depending upon the way in which the cohorts are constructed, the bias in the standard within estimator may still be substantial, even if cohort sizes are fairly large. In general, it holds that, for given n_c, the bias is smaller if the cohorts are chosen such that the relative magnitude of the measurement errors is smaller compared to the within cohort variance of x_{ct}. In practice, however, it may not be easy to construct cohorts in such a way. More recently, Devereux (2007) argues that cell sizes should be much larger, possibly 2000 or more.

In addition to the sizes of the cohorts, the way in which the cohorts are constructed is important. In general, one should be equally careful in choosing cohorts as in selecting instruments. 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.[4] Identification of the parameters in the model requires that the reduced forms in (7) generate sufficient variation over time. This requirement puts a heavy burden on the cohort identifying variables. In particular, it requires that groups

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

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. In that case, the true population cohort means x_{ct} would be identical for each cohort c (and equal the overall population mean) and the only source of variation left in the data that is not attributable to measurement error would be the variation of x_{ct} over time. If these population means do not change over time, all variation in the observed cohort averages $\bar{x}_{ct}$ is measurement error and the errors-in-variables estimator $\hat{\beta}_D$ does not have a well-defined probability limit.

3 Estimation of a linear dynamic model

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 (1) given by

$$y_{it} = \gamma y_{i,t-1} + x'_{it}\beta + \alpha_i + u_{it},\ t = 1, ..., T, \tag{14}$$

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 of Anderson and Hsiao (1981) or, more efficiently, using the GMM estimator of Arellano and Bond (1991). These estimators are based on first-differencing (14) 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 $t-1$. To do so, let z_i denote a set of time-invariant variables, including an intercept term. Now, consider the orthogonal projection in cross-section t of y_{it} upon z_i,

$$E^*\{y_{it}|z_i\} = z'_i\delta_{0t},\ t = 1, ..., T, \tag{15}$$

where E^* denotes the orthogonal projection (for a given t). This is similar to the reduced forms for $x_{k,it}$ in (7). Following Moffitt (1993), one obtains

an estimate of $y_{i,t-1}$ as the predicted value from this regression, substituting the appropriate z values for the individuals in cross-section t. That is,

$$\hat{y}_{i,t-1} = z_i'\hat{\delta}_{0,t-1}, \tag{16}$$

noting that $\hat{\delta}_{0,t-1}$ is estimated from data on different individuals than those indexed by i. In many circumstances it is convenient to think of z_i as a vector of dummy variables, corresponding to mutually exclusive cohorts, as in the previous section. In this case, the orthogonal projection in (15) corresponds to the conditional expectation and (16) corresponds to taking period-by-period sample averages within person i's cohort.

Now, insert these predicted values into the original model to get:

$$y_{it} = \gamma\hat{y}_{i,t-1} + x_{it}'\beta + \varepsilon_{i,t},\ t = 1, ..., T; \tag{17}$$

where

$$\varepsilon_{it} = \alpha_i + u_{it} + \gamma(y_{i,t-1} - \hat{y}_{i,t-1}). \tag{18}$$

No matter how $\hat{y}_{i,t-1}$ is generated, its inclusion implies that one of the explanatory variables is measured with error, although the measurement error will be (asymptotically) uncorrelated with the predicted value.[5] To see whether it would be useful to estimate (17) by ordinary least squares, let us first of all make the assumption that the instruments in z_i are valid, so that

$$E\{(\alpha_i + u_{it})z_i\} = 0,\ t = 1, ..., T. \tag{19}$$

This excludes the possibility that there are cohort effects in the unobservables. While this may appear unreasonable, this assumption is made in Moffitt (1993), Girma (2000) and in a number of cases in McKenzie (2004). Under (19) it can be argued that $\hat{y}_{i,t-1}$ and $\varepsilon_{i,t}$ are uncorrelated, which is a necessary condition for OLS applied to (17) to be consistent. In addition, consistency of OLS requires that x_{it} and ε_{it} are uncorrelated. This assumption may also be problematic, even in cases where the explanatory variables are exogenous to begin with, i.e. even if

$$E\{(\alpha_i + u_{it})x_{it}\} = 0,\ t = 1, ..., T. \tag{20}$$

This is because x_{it} is likely to be correlated with $y_{i,t-1} - \hat{y}_{i,t-1}$.

Consider, for example, a case where high x-values in one period on average correspond with high x-values in the next period. If the β coefficients

[5] Unlike the standard textbook measurement error examples.

are positive this will generally imply that a high value for $x_{i,t-1}$, which is unobservable, will result in an underprediction of $y_{i,t-1}$. On the other hand, $x_{i,t-1}$ is positively correlated with x_{it}. Consequently, this will produce a positive correlation between ε_{it} and x_{it}, resulting in an inconsistent estimator for β. This inconsistency carries over to γ unless $\hat{y}_{i,t-1}$ is uncorrelated with x_{it}. As a result, the estimator suggested by Moffitt (1993), based on applying OLS to (17), is typically inconsistent unless there are either no time-varying exogenous regressors or the time-varying exogenous variables do not exhibit any serial correlation (see Verbeek and Vella, 2005).

To overcome the problem of correlation between the regressors and the error term in (17) one may employ an instrumental variables approach. Note that now we need instruments for x_{it} even though these variables are exogenous in the original model. Because these instruments will have to satisfy a condition like (19), a natural choice is to use the same instruments for x_{it} as we did for $y_{i,t-1}$. This will also guarantee that the instruments are uncorrelated with the prediction error $y_{i,t-1} - \hat{y}_{i,t-1}$ in ε_{it}.

As before, when the instruments z_i are a set of cohort dummies, estimation of (17) by instrumental variables is identical to applying OLS to the original model where all variables are replaced by their (time-specific) cohort sample averages. We can write this as

$$\bar{y}_{ct} = \gamma \bar{y}_{c,t-1} + \bar{x}'_{ct}\beta + \bar{\varepsilon}_{ct}, \;\; c = 1, ..., C; \;\; t = 1, ..., T, \tag{21}$$

where all variables denote period-by-period averages within each cohort. For this approach to be appropriate, we need that $\bar{y}_{c,t-1}$ and $\bar{x}_{ct}$ are not collinear, which requires that the instruments capture variation in $y_{i,t-1}$ independent of the variation in x_{it}. That is, the time-invariant instruments in z_i should exhibit sufficient correlation with the exogenous variables in x_{it} and the (unobserved) lagged dependent variable $y_{i,t-1}$, while at the same time they should not be correlated with ε_{it}. Given these stringent requirements, it is likely that in many applications the number of available valid instruments is small. Verbeek and Vella (2005) provide more details on this rank condition.

The pairwise quasi-differencing approach of Girma (2000) deviates from the above estimation strategy in two respects, although it essentially makes the same assumptions. First, the lagged value of y is not approximated by the lagged cohort average but by an arbitrarily selected observation from the cohort. Second, the instruments are not the cohort dummies, but individual, or averaged, observations from the cohort. As a result, Girma's approach employs a noisy approximation to the unobserved lagged values as well as noisy instruments. Although, under appropriate assumptions, this noise will cancel out asymptotically, there does not seem to be any gain in using such an approach (see Verbeek and Vella, 2005, for more discussion).

The availability of appropriate instruments satisfying condition (19) may be rather limited, because cohort effects in the unobservables are not allowed. It is possible to include cohort fixed effects in essentially the same way as in the static linear model by including the cohort dummies z_i in the equation of interest, with time-invariant coefficients. This imposes (5) and results in

$$y_{it} = \gamma \hat{y}_{i,t-1} + x_{it}'\beta + z_i'\alpha + \eta_{it}, \ t = 1, ..., T, \tag{22}$$

where

$$\eta_{it} = v_i + u_{it} + \gamma(y_{i,t-1} - \hat{y}_{i,t-1}), \tag{23}$$

and $E\{z_i v_i\} = 0$ by construction. This also allows us to relax (20) to

$$E\{(v_i + u_{it})x_{it}\} = 0, \ t = 1, ..., T. \tag{24}$$

Under these conditions, one would estimate (22) by instrumental variables using z_i, interacted with time dummies, as instruments. Verbeek and Vella (2005) refer to this as the augmented IV estimator noting that a time-varying α would make the model unidentified. To achieve identification, we need to assume that $\bar{y}_{c,t-1}$ and $\bar{x}_{ct}$ exhibit time variation and are not collinear. This condition puts additional restrictions upon the relationships between the instruments z_i and x_{it} and $y_{i,t-1}$. Among other things, at least three cross-sections are needed to identify the model under these assumptions.

Computation of this augmented IV estimator is remarkably simple if z_i is a set of cohort dummies. One simply aggregates the data into cohort averages, which gives

$$\bar{y}_{ct} = \gamma \bar{y}_{c,t-1} + \bar{x}_{ct}'\beta + \alpha_c + \bar{\eta}_{c,t}, \tag{25}$$

where $\alpha_c = z_i'\alpha$ denotes a cohort-specific fixed effect. Applying OLS to (25) corresponds to the standard within estimator for $(\gamma, \beta')'$ based upon treating the cohort-level data as a panel, which is consistent under the given assumptions (and some regularity conditions) under type 1 asymptotics ($N \to \infty$ with C fixed). The usual problem with estimating dynamic panel data models (see Nickell, 1981)[6], does not arise because under assumption (24) the error term, which is a within cohort average of individual error terms that are uncorrelated with z_i, is asymptotically zero.[7] However, it remains whether suitable instruments can be found that satisfy the above conditions, because

[6] With genuine panel data, the within estimator in the dynamic model has a substantial bias for small and moderate values of T.

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

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 η_{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. Under type 2 asymptotics ($N \to \infty$ with $C \to \infty$), we encounter similar problems as in the static case, and Collado (1997) discusses how this is handled in the dynamic case, by extending the approach of Deaton (1985). The resulting estimator is similar to the GMM-type estimators that are applied with genuine panel data (Arellano and Bond, 1991), but where the moment matrices are adjusted to reflect the errors-in-variables problem (for finite n_c).

Both Girma (2000) and McKenzie (2004) consider the linear dynamic model with cohort-specific coefficients in equation (14). 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 analyze 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 (14) by having cohort-specific coefficients puts an additional burden upon the identifying conditions. Further, note that using type 2 asymptotics, where the number of cohorts increases with sample size, does not make much sense in these cases.

4 Estimation of a binary choice model

In this section we briefly consider the estimation of a binary choice model on the basis of repeated cross-sections. In a binary choice model the outcome variable takes on only two different values, coded as 0 and 1. For example, the dependent variable could reflect whether or not a household owns a house, or whether or not an individual has a paid job. The model of interest is given by

$$y_{it}^* = x_{it}'\beta + \alpha_i + u_{it}, \ t = 1, ..., T, \tag{26}$$

where y_{it}^* is a latent variable, and we observe

$$\begin{aligned} y_{it} &= 1 \text{ if } y_{it}^* > 0, \\ &= 0 \text{ otherwise.} \end{aligned} \tag{27}$$

With genuine panel data, popular parametric estimators for this model are the random effects probit estimator and the fixed effects logit estimator. The first approach assumes that the unobservables α_i and u_{it} are normally distributed and independent of the explanatory variables in x_{it}. The corresponding likelihood function takes into account that different observations on the same individual are dependent. With repeated cross-sections, this dependence is zero by construction and the binary choice probit model can be estimated as a pooled probit assuming $\alpha_i + u_{it}$ is $N(0,1)$.

Estimation becomes more complicated if one wants to allow α_i and x_{it} to be correlated, as in the fixed effects case. With genuine panel data, one option is to explicitly model this correlation, as in the Chamberlain (1984) approach, who proposes to parametrize the conditional expectation of α_i given the exogenous variables as a linear function of the x_{it}'s. That is,

$$E(\alpha_i | x_{i1}, ..., x_{iT}) = x_{i1}'\lambda_1 + ... + x_{iT}'\lambda_T, \tag{28}$$

which allows us to write

$$\alpha_i = x_{i1}'\lambda_1 + ... + x_{iT}'\lambda_T + \xi_i, \tag{29}$$

where $E(\xi_i | x_{i1}, ..., x_{iT}) = 0$. Substituting (29) into (26) produces

$$y_{it}^* = x_{i1}'\pi_{t1} + ... + x_{iT}'\pi_{tT} + \xi_i + u_{it}, \ t = 1, ..., T, \tag{30}$$

where $\pi_{ts} = \beta + \lambda_s$ if $s = t$ and $\pi_{ts} = \lambda_s$ otherwise. Making distributional assumptions on ξ_i and u_{it} (e.g. normality) allows the application of standard maximum likelihood. However, when only repeated cross-sections are available, we do not observe the full history of the explanatory variables, as required in (29), and this approach is not feasible. Collado (1998) shows how this model can be estimated using cohort data, based on substituting the cohort specific means $\bar{x}_{c1}, ..., \bar{x}_{cT}$ into (30). Using type 2 asymptotics, with $C \to \infty$ and more or less fixed cohort sizes, this introduces an errors-in-variables problem in the equation. However, under normality the covariances between the explanatory variables and the disturbances are known functions of the variances of the measurement error (which can be identified from the individual data). Collado (1998) derives the corresponding probability that $y_{it} = 1$, which can be used to estimate π_t for each cross section t. Next, the

structural parameters β (and λ) can be estimated using a minimum distance estimator. Note that y_{it}^* as well as y_{it} are not aggregated to cohort averages in this approach.

An alternative approach is proposed by Moffitt (1993) and is based on estimating the binary choice model by instrumental variables, where the cohort dummies (or other functions of the variables that define cohorts) are used as instruments. As before, this is based on type 1 asymptotics (with C fixed and $N \to \infty$). Using (5), write the latent variable equation as

$$y_{it}^* = x_{it}'\beta + z_i'\alpha + v_i + u_{it}, \ t = 1, ..., T. \tag{31}$$

Assuming, as before, that the cohort indicators, interacted with time, provide valid instruments, we can estimate the binary choice model by instrumental variables. This requires the assumption that $v_i + u_{it}$ is normally distributed; see Ridder and Moffitt (2007) for more details. Moffitt (1993) and Ridder and Moffitt (2007) also discuss extensions to discrete choice models with a lagged dependent variable.

5 Concluding remarks

In this chapter we have briefly discussed the problem of estimating panel data models from a time series of independent cross-sections. In particular, attention was paid to the estimation of static fixed effects models, to dynamic models with individual effects and to binary choice models.

The approach proposed by Deaton (1985) is to divide the population into a number of cohorts, being groups of individuals sharing some common characteristics, and to treat the observed cohort means as error-ridden measurements of the population cohort means. The resulting estimator for the static linear model with fixed effects is a corrected within estimator based on the cohort aggregates. Moffitt (1993) extends the work of Deaton by considering a general instrumental variables framework, of which explicit grouping is a special kind. While both approaches assume that $N \to \infty$ they differ in the assumptions about what happens to the cohorts when N increases. In Deaton's approach, the number of cohorts C increases with N (with more or less constant cohort sizes), while in Moffitt's approach, the number of cohorts (which is equivalent to the number of instruments) is fixed and cohort sizes increase with N. In this latter approach, the errors-in-variables problem disappears.

Both Moffitt (1993) and Collado (1997) consider the linear dynamic model, based on different types of asymptotics. As argued by Verbeek and Vella

(2005), the fixed effects estimator based on the pseudo panel of cohort averages may provide an attractive choice, even when a lagged dependent variable is included in the model. This deviates from the genuine panel data case, where the standard fixed effects estimator suffers from a substantial small-T bias in dynamic models. A Monte Carlo experiment by Verbeek and Vella (2005) shows that the bias that is present in the within estimator for the dynamic model using genuine panel data (see Nickell, 1981), is much larger than what is found for similar estimators employed upon cohort aggregates.

However, an important issue in both the static and dynamic models is the validity and relevance of the instruments that are used to construct the cohorts. A necessary condition for consistency of most estimators is that all exogenous variables exhibit genuine time-varying cohort-specific variation. That is, the cohorts have exogenous variables that change differentially over time. While it is not obvious that this requirement will be satisfied in empirical applications, it is also not easy to check, because estimation error in the reduced form parameters may hide collinearity problems. That is, sample cohort averages may exhibit time-variation while the unobserved population cohort averages do not.

References

[1] Alessie, R., M.P. Devereux and G. Weber, 1997, Intertemporal Consumption, Durables and Liquidity Constraints: A Cohort Analysis, *European Economic Review*, 41, 37-59.

[2] Anderson, T.W. and C. Hsiao, 1981, Estimation of Dynamic Models with Error Components, *Journal of the American Statistical Association*, 76, 598-606.

[3] Arellano, M. and S. Bond, 1991, Some Test of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, *Review of Economic Studies*, 58, 277-297.

[4] Banks, J., R. Blundell and I. Preston (1994), Life-Cycle Expenditure Allocations and the Consumption Costs of Children, *European Economic Review,* 38, 1391-1410.

[5] Blundell, R., M. Browning, and C. Meghir (1994), Consumer Demand and the Life-Cycle Allocation of Household Expenditures, *Review of Economic Studies*, 61, 57-80.

[6] Blundell, R. A. Duncan and C. Meghir (1998), Estimating Labor Supply Responses Using Tax Reforms, *Econometrica*, 66, 827-861.

[7] Bound, J., D. Jaeger, and R. Baker (1995), Problems with Instrumental Variables Estimation when the Correlation between Instruments and the Endogenous Explanatory Variable is Weak, *Journal of the American Statistical Association*, 90, 443-450.

[8] Browning, M., A. Deaton, and M. Irish (1985), A Profitable Approach to Labor Supply and Commodity Demands over the Life Cycle, *Econometrica*, 53, 503-543.

[9] Chamberlain, G. (1984), Panel Data, in: Z. Griliches and M.D. Intrilligator, eds, *Handbook of Econometrics, Volume 2*, North-Holland, Elsevier Science.

[10] Collado, M.D. (1997), Estimating Dynamic Models from Time Series of Independent Cross-Sections, *Journal of Econometrics*, 82, 37-62.

[11] Collado, M.D. (1998), Estimating Binary Choice Models from Cohort Data, *Investigaciones Económicas*, 22, 259-276.

[12] Deaton, A. (1985), Panel Data from Time Series of Cross Sections, *Journal of Econometrics*, 30, 109-126.

[13] Devereux, P. (2007), Small Sample Bias in Synthetic Cohort Models of Labor Supply, *Journal of Applied Econometrics*, 22, 839–848.

[14] Fuller, W.A. (1987), *Measurement Error Models*, John Wiley and Sons, New York.

[15] Girma, S., 2000, A Quasi-Differencing Approach to Dynamic Modelling from a Time Series of Independent Cross-Sections, *Journal of Econometrics*, 98, 365-383.

[16] Heckman, J.J. and Th.E. McCurdy (1980), A Life-Cycle Model of Female Labour Supply, *Review of Economic Studies*, 47, 47-74.

[17] McKenzie, D.J. (2004), Asymptotic Theory for Heterogeneous Dynamic Pseudo-Panels, *Journal of Econometrics,* 120, 235-262.

[18] Moffitt, R. (1993), Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-Sections, *Journal of Econometrics*, 59, 99-123.

[19] Nickell, S. (1981), Biases in Dynamic Models with Fixed Effects, *Econometrica*, 49, 1417-1426.

[20] Pesaran, M.H. and R. Smith (1995), Estimating Long-Run Relationships from Dynamic Heterogeneous Panels, *Journal of Econometrics*, 68, 79-113.

[21] Propper, C., H. Rees and K. Green (2001), The Demand for Private Medical Insurance in the UK: A Cohort Analysis, *The Economic Journal*, 111, C180-C200.

[22] Ridder, G, and R. Moffitt (2007), *The Econometrics of Data Combination*, in: J.J. Heckman and E.E. Leamer, eds., *Handbook of Econometrics, Volume 6B*, North-Holland, Elsevier Science.

[23] Robertson, D. and J. Symons (1992), Some Strange Properties of Panel Data Estimators, *Journal of Applied Econometrics*, 7, 175-189.

[24] Staiger, D. and J.H. Stock (1997), Instrumental Variables Regressions with Weak Instruments, *Econometrica*, 65, 557-586.

[25] Verbeek, M. and Th.E. Nijman (1992), Can Cohort Data Be Treated As Genuine Panel Data?, *Empirical Economics*, 17, 9-23.

[26] Verbeek, M. and Th.E. Nijman (1993), Minimum MSE Estimation of a Regression Model with Fixed Effects from a Series of Cross-Sections, *Journal of Econometrics*, 59, 125-136.

[27] Verbeek, M. and F. Vella (2005), Estimating Dynamic Models from Repeated Cross-Sections, *Journal of Econometrics*, 127, 83-102.