Chapter 53

PANEL DATA MODELS: SOME RECENT DEVELOPMENTS*

MANUEL ARELLANO

CEMFI, Casado del Alisal 5, 28014 Madrid, Spain

BO HONORÉ

Department of Economics, Princeton University, Princeton, New Jersey 08544

Econ 312, Spring 2021

* We thank Jason Abrevaya, Badi Baltagi, Olympia Bover, Martin Browning, Jim Heckman, Luojia Hu, Ekaterini Kyriazidou, Ed Leamer, Aprajit Mahajan, Enrique Sentana, Jeffrey Wooldridge and participants at the Chicago and London Handbook conferences for helpful comments. All errors are our responsibility.

Handbook of Econometrics, Volume 5, Edited by J.J. Heckman and E. Leamer

Abstract

This chapter focuses on two of the developments in panel data econometrics since the Handbook chapter by Chamberlain (1984).

The first objective of this chapter is to provide a review of linear panel data models with predetermined variables. We discuss the implications of assuming that explanatory variables are predetermined as opposed to strictly exogenous in dynamic structural equations with unobserved heterogeneity. We compare the identification from moment conditions in each case, and the implications of alternative feedback schemes for the time series properties of the errors. We next consider autoregressive error component models under various auxiliary assumptions. There is a trade-off between robustness and efficiency since assumptions of stationary initial conditions or time series homoskedasticity can be very informative, but estimators are not robust to their violation. We also discuss the identification problems that arise in models with predetermined variables and multiple effects. Concerning inference in linear models with predetermined variables, we discuss the form of optimal instruments, and the sampling properties of GMM and LIML-analogue estimators drawing on Monte Carlo results and asymptotic approximations.

A number of identification results for limited dependent variable models with fixed effects and strictly exogenous variables are available in the literature, as well as some results on consistent and asymptotically normal estimation of such models. There are also some results available for models of this type including lags of the dependent variable, although even less is known for nonlinear dynamic models. Reviewing the recent work on discrete choice and selectivity models with fixed effects is the second objective of this chapter. A feature of parametric limited dependent variable models is their fragility to auxiliary distributional assumptions. This situation prompted the development of a large literature dealing with semiparametric alternatives (reviewed in Powell, 1994's chapter). The work that we review in the second part of the chapter is thus at the intersection of the panel data literature and that on cross-sectional semiparametric limited dependent variable models.

1. Introduction

Panel data analysis is at the watershed of time series and cross-section econometrics. While the identification of time series parameters traditionally relied on notions of stationarity, predeterminedness and uncorrelated shocks, cross-sectional parameters appealed to exogenous instrumental variables and random sampling for identification. By combining the time series and cross-sectional dimensions, panel datasets have enriched the set of possible identification arrangements, and forced economists to think more carefully about the nature and sources of identification of parameters of potential interest.

One strand of the literature found its original motivation in the desire of exploiting panel data for controlling unobserved time-invariant heterogeneity in cross-sectional models. Another strand was interested in panel data as a way to disentangle components of variance and to estimate transition probabilities among states. Papers in these two veins can be loosely associated with the early work on fixed and random effects approaches, respectively. In the former, interest typically centers in measuring the effect of regressors holding unobserved heterogeneity constant. In the latter, the parameters of interest are those characterizing the distributions of the error components. A third strand of the literature studied autoregressive models with individual effects, and more generally models with lagged dependent variables.

A sizeable part of the work in the first two traditions concentrated on models with just strictly exogenous variables. This contrasts with the situation in time series econometrics where the distinction between predetermined and strictly exogenous variables has long been recognized as a fundamental one in the specification of empirical models.

The first objective of this chapter is to review recent work on linear panel data models with predetermined variables. Lack of control of individual heterogeneity could result in a *spurious* rejection of strict exogeneity, and so a definition of strict exogeneity conditional on unobserved individual effects is a useful extension of the standard concept to panel data (a major theme of Chamberlain, 1984's chapter). There are many instances, however, in which for theoretical or empirical reasons one is concerned with models exhibiting *genuine* lack of strict exogeneity after controlling for individual heterogeneity.

The interaction between unobserved heterogeneity and predetermined regressors in short panels – which are the typical ones in microeconometrics – poses identification problems that are absent from both time series models and panel data models with only strictly exogenous variables. In our review we shall see that for linear models it is possible to accommodate techniques developed from the various strands in a common framework within which their relative merits can be evaluated.

Much less is known for discrete choice, selectivity and other non-linear models of interest in microeconometrics. A number of identification results for limited dependent variable models with fixed effects and strictly exogenous variables are available in the literature, as well as some results on consistent and asymptotically normal estimation of

such models. There are also some results available for models of this type including lags of the dependent variable, although even less is known for nonlinear dynamic models.

Reviewing the recent work on discrete choice and selectivity models with fixed effects is the second objective of this chapter. A feature of parametric limited dependent variable models is their fragility to auxiliary distributional assumptions. This situation prompted the development of a large literature dealing with semiparametric alternatives (reviewed in Powell, 1994's chapter). The work that we review in the second part of the chapter is thus at the intersection of the panel data literature and that on cross-sectional semiparametric limited dependent variable models.

Other interesting topics in panel data analysis which will not be covered in this chapter include work on long T panel data models with heterogeneous dynamics or unit roots [Pesaran and Smith (1995), Canova and Marcet (1995), Kao (1999), Phillips and Moon (1999)], simulation-based random effects approaches to the nonlinear models [Hajivassiliou and McFadden (1990), Keane (1993, 1994), Allenby and Rossi (1999), and references therein], classical and Bayesian flexible estimators of error component distributions [Horowitz and Markatou (1996), Chamberlain and Hirano (1999), Geweke and Keane (2000)], other nonparametric and semiparametric panel data models [Baltagi, Hidalgo and Li (1996), Li and Stengos (1996), Li and Hsiao (1998) and Chen, Heckman and Vytlacil (1998)], and models from time series of independent cross-sections [Deaton (1985), Moffitt (1993), Collado (1997)]. Some of these topics as well as comprehensive reviews of the panel data literature are covered in the text books by Hsiao (1986) and Baltagi (1995).

2. Linear models with predetermined variables: identification

In this section we discuss the identification of linear models with predetermined variables in two different contexts. In Section 2.1 the interest is to identify structural parameters in models in which explanatory variables are correlated with a time-invariant individual effect, but they are either strictly exogenous or predetermined relative to the time-varying errors. The second context, discussed in Section 2.2, is the time series analysis of error component models with autoregressive errors under various auxiliary assumptions. Section 2.3 discusses the use of stationarity restrictions in regression models, and Section 2.4 considers the identification of models with multiplicative or multiple individual effects.

2.1. Strict exogeneity, predeterminedness, and unobserved heterogeneity

We begin with a discussion of the implications of strict exogeneity for identification of regression parameters controlling for unobserved heterogeneity, with the objective of comparing this situation with that where the regressors are only predetermined variables.

Static regression with a strictly exogenous variable. Let us consider a linear regression for panel data including a fixed effect η_i and a time effect δ_t with N individuals observed T time periods, where T is small and N is large:

$$y_{it} = \beta x_{it} + \delta_t + \eta_i + \upsilon_{it} \quad (i = 1, \ldots, N;\ t = 1, \ldots, T). \tag{1}$$

We assume that $(y_{i1} \cdots y_{iT}, x_{i1} \cdots x_{iT}, \eta_i)$ is an iid random vector with finite second-order moments, while β and the time effects are treated as unknown parameters. The variable x_{it} is said to be strictly exogenous in this model if it is uncorrelated with past, present and future values of the disturbance υ_{it}:

$$E^*(\upsilon_{it} | x_i^T) = 0 \quad (t = 1, \ldots, T), \tag{2}$$

where E^* denotes a linear projection, and we use the superscript notation $z_i^t = (z_{i1}, \ldots, z_{it})'$. First-differencing the conditions we obtain

$$E^*(\upsilon_{it} - \upsilon_{i(t-1)} | x_i^T) = 0 \quad (t = 2, \ldots, T). \tag{3}$$

Since in the absence of any knowledge about η_i the condition $E^*(v_{i1}|x_i^T) = 0$ is not informative about β, the restrictions in first-differences are equivalent to those in levels. Therefore, for fixed T the problem of cross-sectional identification of β is simply that of a multivariate regression in first differences subject to cross-equation restrictions, and β is identifiable with $T \geqslant 2$.

Specifically, letting $E^*(\eta_i|x_i^T) = \lambda_0 + \lambda' x_i^T$, the model can be written as

$$y_{it} = \pi_{0t} + \beta x_{it} + \lambda' x_i^T + \varepsilon_{it} \text{ with } E^*(\varepsilon_{it}|x_i^T) = 0 \quad (t = 1, \ldots, T). \tag{4}$$

where $\pi_{0t} = \lambda_0 + \delta_t$. This T equation system is equivalent to

$$y_{i1} = \pi_{01} + \beta x_{i1} + \lambda' x_i^T + \varepsilon_{i1} \qquad E^*(\varepsilon_{i1}|x_i^T) = 0, \tag{5}$$

$$\Delta y_{it} = \Delta\delta_t + \beta \Delta x_{it} + \Delta\varepsilon_{it} \qquad E^*(\Delta\varepsilon_{it}|x_i^T) = 0 \quad (t = 2, \ldots, T). \tag{6}$$

In the absence of restrictions in λ Equation (5) is uninformative about β, and as a consequence asking under which conditions β is identified in Equation (4) is equivalent to asking under which conditions β is identified in Equation (6)[1].

[1] Lack of dependence between v_{it} and x_i^T could also be expressed in terms of conditional independence in mean $E(v_{it}|x_i^T) = 0$ $(t = 1, \ldots, T)$. In the absence of any knowledge about η_i this is equivalent to the $(T-1)$ conditional moment restrictions $E(v_{it} - v_{i(t-1)}|x_i^T) = 0$ $(t = 2, \ldots, T)$ which do not depend on η_i [Chamberlain (1992a)]. In the presentation for linear models, however, the use of linear projections affords a straightforward discussion of identification, and in the context of estimation it allows us to abstract from issues relating to optimal instruments and semiparametric asymptotic efficiency.

Partial adjustment with a strictly exogenous variable. In an alternative model, the effect of a strictly exogenous x on y could be specified as a partial adjustment equation:

$$y_{it} = \alpha y_{i(t-1)} + \beta_0 x_{it} + \beta_1 x_{i(t-1)} + \delta_t + \eta_i + \upsilon_{it} \quad (i = 1, \ldots, N;\ t = 2, \ldots, T) \qquad (7)$$

together with

$$E^*(\upsilon_{it} | x_i^T) = 0 \quad (t = 2, \ldots, T). \qquad (8)$$

Note that assumption (8) does not restrict the serial correlation of υ, so that lagged y is an endogenous explanatory variable. In the equation in levels, $y_{i(t-1)}$ will be correlated with η_i by construction and may also be correlated with past, present and future values of the errors υ_{it} since they may be autocorrelated in an unspecified way. Likewise, the system in first differences is free from fixed effects and satifies $E^*(\Delta\upsilon_{it} | x_i^T) = 0$ $(t = 3, \ldots, T)$, but $\Delta y_{i(t-1)}$ may still be correlated with $\Delta\upsilon_{is}$ for all s.

Subject to a standard rank condition, α, β_0, β_1 and the time effects will be identified with $T \geqslant 3$. With $T = 3$ they are just identified since there are five orthogonality conditions and five unknown parameters:

$$E\left[\begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} (\Delta y_{i3} - \alpha \Delta y_{i2} - \beta_0 \Delta x_{i3} - \beta_1 \Delta x_{i2} - \Delta \delta_3)\right] = 0. \tag{9}$$

$$E\,(y_{i2} - \alpha y_{i1} - \beta_0 x_{i2} - \beta_1 x_{i1} - \delta_2) = 0.$$

This simple example illustrates the potential for cross-sectional identification under strict exogeneity. In effect, strict exogeneity of x permits the identification of the dynamic effect of x on y and of lagged y on current y, in the presence of a fixed effect and shocks that can be arbitrarily persistent over time [cf. Bhargava and Sargan (1983), Chamberlain (1982a, 1984), Arellano (1990)].

A related situation of economic interest arises in testing life-cycle models of consumption or labor supply with habits [e.g., Bover (1991), or Becker, Grossman and Murphy (1994)]. In these models the coefficient on the lagged dependent variable is a parameter of central interest as it is intended to measure the extent of habits. However, in the absence of an exogenous instrumental variable such a coefficient would not be identified, since the effect of genuine habits could not be separated from serial correlation in the unobservables.

As an illustration, let us consider the empirical model of cigarette consumption by Becker, Grossman and Murphy (1994) for US state panel data. Their empirical analysis is based on the following equation:

$$c_{it} = \theta c_{i\,(t-1)} + \beta\theta c_{i\,(t+1)} + \gamma p_{it} + \eta_i + \delta_t + \upsilon_{i\,(t+1)}, \tag{10}$$

where c_{it} and p_{it} denote, respectively, annual per capita cigarette consumption in packs by state and average cigarette price per pack. Becker et al. are interested in testing

whether smoking is addictive by considering the response of cigarette consumption to a change in cigarette prices.

The rationale for Equation (10) is provided by a model of addictive behavior in which utility in period t depends on cigarette consumption in t and in $t-1$. Under perfect certainty and quadratic utility, the equation can be obtained from the first-order conditions of utility maximization. The degree of addiction is measured by θ, which will be positive if smoking is addictive. The current price coefficient γ should be negative by concavity of the utility, and β denotes the discount factor. With certainty, the marginal utility of wealth is constant over time but not cross-sectionally. The state specific intercept η_i is meant to capture such variation[2]. Finally, the δ_t's represent aggregate shocks, possibly correlated with prices, which are treated as period specific parameters.

The errors $v_{i(t+1)}$ capture unobserved life-cycle utility shifters, which are likely to be serially correlated. Therefore, even in the absence of addiction ($\theta = 0$) and serial correlation in prices, we would expect c_{it} to be autocorrelated, and in particular to find a non-zero effect of $c_{i(t-1)}$ in a linear regression of c_{it} on $c_{i(t-1)}$, $c_{i(t+1)}$ and p_{it}. Current consumption depends on prices in all periods through the effects of past and future consumption, but it is independent of past and future prices when $c_{i(t-1)}$ and $c_{i(t+1)}$ are held fixed. Thus, Becker et al.'s strategy is to identify θ, β, and γ from the assumption that prices are strictly exogenous relative to the unobserved utility shift variables. The required exogenous variation in prices comes from the variation in cigarette tax rates across states and time, and agents are assumed to be able to anticipate future prices without error.

Partial adjustment with a predetermined variable. The assumption that current values of x are not influenced by past values of y and v is often unrealistic. We shall say that x is predetermined in a model like Equation (7) if

$$E^*(v_{it}|x_i^t, y_i^{t-1}) = 0 \quad (t = 2, \ldots, T). \tag{11}$$

That is, current shocks are uncorrelated with past values of y and with current and past values of x, but feedback effects from lagged dependent variables (or lagged errors) to current and future values of the explanatory variable are not ruled out.

Note that, in contrast with Equation (8), assumption (11) does restrict the serial correlation of υ. Specifically, it implies that the errors in first differences exhibit first-order autocorrelation but are uncorrelated at all other lags:

$$E(\Delta\upsilon_{it}\Delta\upsilon_{i(t-j)}) = 0 \quad j > 1.$$

Examples of this situation include Euler equations for household consumption [Zeldes (1989), Runkle (1991), Keane and Runkle (1992)], or for company investment

[2] According to the theory γ would also be state specific, since it is a function of the marginal utility of wealth. Thus the model with constant price coefficient must be viewed as an approximate model.

[Bond and Meghir (1994)], in which variables in the agents' information sets are uncorrelated with current and future idiosyncratic shocks but not with past shocks, together with the assumption that the empirical model's errors are given by such shocks.

Another example is the effect of children on female labour force participation decisions. In this context, assuming that children are strictly exogenous is much stronger than the assumption of predeterminedness, since it would require us to maintain that labour supply plans have no effect on fertility decisions at any point in the life cycle [Browning (1992, p. 1462)].

The implication of Equation (11) for errors in first differences is that

$$E^*(v_{it} - v_{i(t-1)} | x_i^{t-1}, y_i^{t-2}) = 0 \quad (t = 3, \ldots, T). \tag{12}$$

As before, these restrictions are equivalent to those in levels since in the absence of any knowledge about η_i the levels are not informative about the parameters[3]. Subject to a rank condition, α, β_0, β_1 and the time effects will be identified with $T \geqslant 3$. With $T = 3$ they are just identified from the five orthogonality conditions:

$$\begin{aligned} E[\begin{pmatrix} 1 \\ y_{i1} \\ x_{i1} \\ x_{i2} \end{pmatrix} (\Delta y_{i3} - \alpha \Delta y_{i2} - \beta_0 \Delta x_{i3} - \beta_1 \Delta x_{i2} - \Delta \delta_3)] &= 0, \\ E(y_{i2} - \alpha y_{i1} - \beta_0 x_{i2} - \beta_1 x_{i1} - \delta_2) &= 0. \end{aligned} \tag{13}$$

It is of some interest to compare the situation in Equation (13) with that in Equation (9). The two models are not nested since they only have four moment restrictions in common, which in this example are not sufficient to identify the five parameters. The model with a strictly exogenous x would become a special case of the model with a predetermined x, only if in the former serial correlation were ruled out. That is, if Equation (8) were replaced with:

$$E^*(v_{it}|x_i^T, y_i^{t-1}) = 0 \quad (t = 2, \ldots, T). \tag{14}$$

However, unlike in the predetermined case, lack of *arbitrary* serial correlation is not an identification condition for the model with strict exogeneity.

In the predetermined case it is still possible to accommodate special forms of serial correlation. For example, with $T = 4$ the parameters in the dynamic model are just identified with $E(\Delta v_{it}\Delta v_{i(t-j)}) = 0$ for $j > 2$, which is consistent with a first-order

[3] Orthogonality conditions of this type have been considered by Anderson and Hsiao (1981, 1982), Griliches and Hausman (1986), Holtz-Eakin, Newey and Rosen (1988), and Arellano and Bond (1991) amongst others.

moving average process for v. This is so because in such case there are still three valid orthogonality restrictions: $E(y_{i1}\Delta v_{i4}) = 0$, $E(x_{i1}\Delta v_{i4}) = 0$, and $E(x_{i2}\Delta v_{i4}) = 0$.

Uncorrelated errors arise as the result of theoretical predictions in a number of environments (e.g., innovations in rational expectation models). However, even in the absence of specific restrictions from theory, the nature of shocks in econometric models is often less at odds with assumptions of no or limited autocorrelation than with the absence of feedback in the explanatory variable processes[4].

In the previous discussion we considered models for which the strict exogeneity property was unaffected by serial correlation, and models with feedback from lagged y or v to current values of x, but other situations are possible. For example, it may be the case that the strict exogeneity condition (2) for model (1) is only satisfied as long as errors are unpredictable. An illustration is the agricultural Cobb–Douglas production function discussed by Chamberlain (1984), where y is log output, x is log labor, η is soil quality, and v is rainfall. If η is known to farmers and they choose x to maximize expected profits, x will be correlated with η, but uncorrelated with v at all lags and leads provided v is unpredictable from past rainfall. If rainfall in t is predictable from rainfall in $t-1$, labour demand in t will in general depend on $v_{i(t-1)}$ [Chamberlain (1984, pp. 1258–1259)].

Another situation of interest is a case where the model is (1) or (7) and we only condition on x_i^t. That is, instead of Equation (11) we have

$$E^*(\upsilon_{it} \mid x_i^t) = 0. \tag{15}$$

In this case serial correlation is not ruled out, and the partial adjustment model is identifiable with $T \geqslant 4$, but Equation (15) rules out unspecified feedback from lagged y to current x. As an example, suppose that $\upsilon_{it} = \zeta_{it} + \varepsilon_{it}$ is an Euler equation's error given by the sum of a serially correlated preference shifter ζ_{it} and a white noise expectation error ε_{it}. The υ's will be serially correlated and correlated with lagged consumption variables y but not with lagged price variables x. Another example is an equation $y_{it}^* = \beta x_{it} + \eta_i + \upsilon_{it}^*$ where υ_{it}^* is white noise and x_{it} depends on $y_{i(t-1)}^*$, but y_{it}^* is measured with an autocorrelated error independent of x and y^* at all lags and leads.

Implications of uncorrelated effects. So far, we have assumed that all the observable variables are correlated with the fixed effect. If a strictly exogenous x were known to be uncorrelated with η, the parameter β in the static regression (1) would be identified from a single cross-section ($T = 1$). However, in the dynamic regression the lagged dependent variable would still be correlated with the effects by construction, so knowledge of lack of correlation between x and η would add T orthogonality conditions to the ones discussed above, but the parameters would still be identified

[4] As an example, see related discussions on the specification of shocks in Q investment equations by Hayashi and Inoue (1991), and Blundell, Bond, Devereux and Schiantarelli (1992).

only when $T \geqslant 3$[5]. The moment conditions for the partial adjustment model with strictly exogenous x and uncorrelated effects can be written as

$$E[\begin{pmatrix} 1 \\ x_i^T \end{pmatrix}(y_{it} - \alpha y_{i(t-1)} - \beta_0 x_{it} - \beta_1 x_{i(t-1)} - \delta_t)] = 0 \quad (t = 2, \ldots, T). \tag{16}$$

A predetermined x could also be known to be uncorrelated with the fixed effects if feedback occurred from lagged errors but not from lagged y. To illustrate this point suppose that the process for x is

$$x_{it} = \rho x_{i(t-1)} + \gamma v_{i(t-1)} + \phi \eta_i + \varepsilon_{it}, \tag{17}$$

where ε_{it}, v_{is} and η_i are mutually uncorrelated for all t and s. In this example x is uncorrelated with η when $\phi = 0$. However, if $v_{i(t-1)}$ were replaced by $y_{i(t-1)}$ in Equation (17), x and η will be correlated in general even with $\phi = 0$. Knowledge of lack of correlation between a predetermined x and η would also add T orthogonality restrictions to the ones discussed above for such a case. The moment conditions for the partial adjustment model with a predetermined x uncorrelated with the effects can be written as

$$E[\begin{pmatrix} 1 \\ x_i^t \end{pmatrix}(y_{it} - \alpha y_{i(t-1)} - \beta_0 x_{it} - \beta_1 x_{i(t-1)} - \delta_t)] = 0 \quad (t = 2, \ldots, T), \tag{18}$$

$$E[y_i^{t-2}(\Delta y_{it} - \alpha \Delta y_{i(t-1)} - \beta_0 \Delta x_{it} - \beta_1 \Delta x_{i(t-1)} - \Delta \delta_t)] = 0 \quad (t = 3, \ldots, T).$$

Again, the parameters in this case would only be identified when $T \geqslant 3$.

Relationship with statistical definitions. To conclude this discussion, it may be useful to relate our usage of strict exogeneity to statistical definitions. A (linear projection based) *statistical* definition of strict exogeneity conditional on a fixed effect would state that x is strictly exogenous relative to y given η if

$$E^*(y_{it}|x_i^T, \eta_i) = E^*(y_{it}|x_i^t, \eta_i). \tag{19}$$

This is equivalent to the statement that y does not Granger-cause x given η in the sense that

$$E^*(x_{i(t+1)}|x_i^t, y_i^t, \eta_i) = E^*(x_{i(t+1)}|x_i^t, \eta_i). \tag{20}$$

Namely, letting $x_i^{(t+1)T} = (x_{i(t+1)}, \ldots, x_{iT})'$ if we have

$$E^*(y_{it}|x_i^T, \eta_i) = \beta_t' x_i^t + \delta_t' x_i^{(t+1)T} + \gamma_t \eta_i \tag{21}$$

and

$$E^*(x_{i(t+1)}|x_i^t, y_i^t, \eta_i) = \psi_t' x_i^t + \phi_t' y_i^t + \varsigma_t \eta_i, \tag{22}$$

it turns out that the restrictions $\delta_t = 0$ and $\phi_t = 0$ are equivalent. This result generalized the well-known equivalence between strict exogeneity [Sims (1972)] and Granger's

[5] Models with strictly exogenous variables uncorrelated with the effects were considered by Hausman and Taylor (1981), Bhargava and Sargan (1983), Amemiya and MaCurdy (1986), Breusch, Mizon and Schmidt (1989), Arellano (1993), and Arellano and Bover (1995).

non-causality [Granger (1969)][6]. It was due to Chamberlain (1984), and motivated the analysis in Holtz-Eakin, Newey and Rosen (1988), which was aimed at testing such a property.

Here, however, we are using strict exogeneity relative to the errors of an econometric model. Strict exogeneity itself, or the lack of it, may be a property of the model suggested by theory. We used some simple models as illustrations, in the understanding that the discussion would also apply to models that may include other features like individual effects uncorrelated with errors, endogenous explanatory variables, autocorrelation, or constraints in the parameters. Thus, in general strict exogeneity relative to a model may or may not be testable, but if so we shall usually be able to test it only in conjunction with other features of the model. In contrast with the econometric concept, a statistical definition of strict exogeneity is model free, but whether it is satisfied or not, may not necessarily be of relevance for the econometric model of interest[7].

As an illustration, let us consider a simple permanent-income model. The observables are non-durable expenditures c_{it}, current income w_{it}, and housing expenditure x_{it}. The unobservables are permanent (w_{it}^p) and transitory (ε_{it}) income, and measurement errors in non-durable (ξ_{it}) and housing (ς_{it}) expenditures. The expenditure variables are assumed to depend on permanent income only, and the unobservables are mutually independent but can be serially correlated. With these assumptions we have

$$w_{it} = w_{it}^p + \varepsilon_{it}, \tag{23}$$

$$c_{it} = \beta w_{it}^p + \xi_{it}, \tag{24}$$

$$x_{it} = \gamma w_{it}^p + \varsigma_{it}. \tag{25}$$

Suppose that β is the parameter of interest. The relationship between c_{it} and w_{it} suggested by the theory is of the form

$$c_{it} = \beta w_{it} + \upsilon_{it}, \tag{26}$$

where $\upsilon_{it} = \xi_{it} - \beta\varepsilon_{it}$. Since w_{it} and υ_{it} are contemporaneously correlated, w_{it} is an endogenous explanatory variable in Equation (26). Moreover, since $E^*(\upsilon_{it}|x_i^T) = 0$, x_{it} is a strictly exogenous instrumental variable in Equation (26). At the same time, note

[6] If linear projections are replaced by conditional distributions, the equivalence does not hold and it turns out that the definition of Sims is weaker than Granger's definition. Conditional Granger non-causality is equivalent to the stronger Sims' condition given by $f(y_t|x^T, y^{t-1}) = f(y_t|x^t, y^{t-1})$ [Chamberlain (1982b)].

[7] Unlike the linear predictor definition, a conditional independence definition of strict exogeneity given an individual effect is not restrictive, in the sense that there always exists a random variable η such that the condition is satisfied [Chamberlain (1984)]. This lack of identification result implies that a conditional-independence test of strict exogeneity given an individual effect will necessarily be a joint test involving a (semi) parametric specification of the conditional distribution.

that in general linear predictors of x given its past can be improved by adding lagged values of c and/or w (unless permanent income is white noise). Thus, the statistical condition for Granger non-causality or strict exogeneity is not satisfied in this example. A similar discussion could be conducted for a version of the model including fixed effects.

2.2. Time series models with error components

The motivation in the previous discussion was the identification of regression responses not contaminated from heterogeneity biases. Another leading motivation for using panel data is the analysis of the time series properties of the observed data. Models of this kind were discussed by Lillard and Willis (1978), MaCurdy (1982), Hall and Mishkin (1982), Holtz-Eakin, Newey and Rosen (1988) and Abowd and Card (1989), amongst others.

An important consideration is distinguishing unobserved heterogeneity from genuine dynamics. For example, the exercises cited above are all concerned with the time series properties of individual earnings for different reasons, including the analysis of earnings mobility, testing the permanent income hypothesis, or estimating intertemporal labour supply elasticities. However, how much dependence is measured in the residuals of the earnings process depends crucially, not only on how much heterogeneity is allowed into the process, but also on the auxiliary assumptions made in the specification of the residual process, and assumptions about measurement errors.

One way of modelling dynamics is through moving average processes [e.g., Abowd and Card (1989)]. These processes limit persistence to a fixed number of periods, and imply linear moment restrictions in the autocovariance matrix of the data. Autoregressive processes, on the other hand, imply nonlinear covariance restrictions but provide instrumental-variable orthogonality conditions that are linear in the autoregressive coefficients. Moreover, they are well suited to analyze the implications for identification and inference of issues such as the stationarity of initial conditions, homoskedasticity, and (near) unit roots.

Another convenient feature of autoregressive processes is that they can be regarded as a special case of the regression models with predetermined variables discussed above. This makes it possible to consider both types of problems in a common framework, and facilitates the distinction between static responses with residual serial correlation and dynamic responses[8]. Finally, autoregressive models are more easily extended to limited-dependent-variable models.

In the next subsection we discuss the implications for identification of alternative assumptions concerning a first-order autoregressive process with individual effects in short panels.

[8] In general, linear conditional models can be represented as data covariance matrix structures, but typically they involve a larger parameter space including many nuisance parameters, which are absent from instrumental-variable orthogonality conditions.

2.2.1. *The AR(1) process with fixed effects*[9]

Let us consider a random sample of individual time series of size T, $\{y_i^T, i = 1, \ldots, N\}$, with second-order moment matrix $E(y_i^T y_i^{T\prime}) = \Omega = \{\omega_{ts}\}$. We assume that the joint distribution of y_i^T and the individual effect η_i satisfies

$$y_{it} = \alpha y_{i(t-1)} + \eta_i + \upsilon_{it} \quad (i = 1, \ldots, N; \quad t = 2, \ldots, T) \quad |\alpha| < 1, \tag{27}$$

$$E^*(\upsilon_{it} | y_i^{t-1}) = 0 \quad (t = 2, \ldots, T), \tag{A1}$$

where $E(\eta_i) = \gamma$, $E(\upsilon_{it}^2) = \sigma_t^2$, and $\mathrm{Var}(\eta_i) = \sigma_\eta^2$. Notice that the assumption does not rule out correlation between η_i and υ_{it}, nor the possibility of conditional heteroskedasticity, since $E(\upsilon_{it}^2 | y_i^{t-1})$ need not coincide with σ_t^2. Equations (27) and (A1) can be seen as a specialization of Equations (7) and (11). Thus, following the discussion above, (A1) implies $(T-2)(T-1)/2$ linear moment restrictions of the form

$$E[\, y_i^{t-2}(\Delta y_{it} - \alpha \Delta y_{i(t-1)})] = 0. \tag{28}$$

These restrictions can also be represented as constraints on the elements of Ω. Multiplying Equation (27) by y_{is} for $s < t$, and taking expectations gives $\omega_{ts} = \alpha\omega_{(t-1)s} + c_s$, $(t = 2, \ldots, T;\ s = 1, \ldots, t-1)$, where $c_s = E(y_{is}\eta_i)$. This means that, given assumption A1, the $T(T+1)/2$ different elements of Ω can be written as functions of the $2T \times 1$ parameter vector $\theta = (\alpha, c_1, \ldots, c_{T-1}, \omega_{11}, \ldots, \omega_{TT})'$. Notice that with $T = 3$ the parameters (α, c_1, c_2) are just identified as functions of the elements of Ω:

$$\begin{aligned}
\alpha &= (\omega_{21} - \omega_{11})^{-1}(\omega_{31} - \omega_{21}) \\
c_1 &= \omega_{21} - \alpha\omega_{11} \\
c_2 &= \omega_{32} - \alpha\omega_{22}.
\end{aligned}$$

The model based on A1 is attractive because the identification of α, which measures persistence given unobserved heterogeneity, is based on minimal assumptions. However, we may be willing to impose additional structure if this conforms to a priori beliefs.

Lack of correlation between the effects and the errors. One possibility is to assume that the errors v_{it} are uncorrelated with the individual effect η_i given y_i^{t-1}. In a structural context, this will often be a reasonable assumption if, for example, the v_{it} are interpreted as innovations that are independent of variables in the agents' information

[9] This section follows a similar discussion by Alonso-Borrego and Arellano (1999).

set. In such case, even if η_i is not observable to the econometrician, being time-invariant it is likely to be known to the individual. This situation gives rise to the following assumption

$$E^*(\upsilon_{it} | y_i^{t-1}, \eta_i) = 0 \quad (t = 2, \ldots, T). \tag{A1'}$$

Note that in a short panel assumption A1′ is more restrictive than assumption A1. Nevertheless, lack of correlation between υ_{it} and $\{y_{i(t-1)}, \ldots, y_{i(t-J)}\}$ implies lack of correlation between υ_{it} and η_i in the limit as $J \to \infty$. This will be so as long as

$$\eta_i = \operatorname*{plim}_{J \to \infty} \frac{1}{J} \sum_{j=1}^{J} \left(y_{i(t-j)} - \alpha y_{i(t-j-1)} \right).$$

Thus, for a process that started at $-\infty$ we would have orthogonality between η_i and υ_{it}, and any correlation between individual effects and shocks will tend to vanish as t increases.

When $T \geqslant 4$, assumption A1′ implies the following additional $T - 3$ quadratic moment restrictions that were considered by Ahn and Schmidt (1995):

$$E[(y_{it} - \alpha y_{i(t-1)})(\Delta y_{i(t-1)} - \alpha \Delta y_{i(t-2)})] = 0 \quad (t = 4, \ldots, T). \tag{29}$$

In effect, we can write $E[(y_{it} - \alpha y_{i(t-1)} - \eta_i)(\Delta y_{i(t-1)} - \alpha \Delta y_{i(t-2)})] = 0$ and since $E(\eta_i \Delta \upsilon_{i(t-1)}) = 0$ the result follows. Thus, Equation (29) also holds if $\mathrm{Cov}(\eta_i, \upsilon_{it})$ is constant over t.

An alternative representation of the restrictions in Equation (29) is in terms of a recursion of the coefficients c_t introduced above. Multiplying Equation (27) by η_i and taking expectations gives $c_t = \alpha c_{t-1} + \phi, (t = 2, \ldots, T)$, where $\phi = E(\eta_i^2) = \gamma^2 + \sigma_\eta^2$, so that $c_1, \ldots, c_T$ can be written in terms of c_1 and ϕ. This gives rise to a covariance structure in which Ω depends on the $(T + 3) \times 1$ parameter vector $\theta = (\alpha, \phi, c_1, \omega_{11}, \ldots, \omega_{TT})'$. Notice that with $T = 3$ assumption A1′ does not imply further restrictions in Ω, with the result that α remains just identified. One can solve for ϕ in terms of α, c_1 and c_2:

$$\phi = (\omega_{32} - \omega_{21}) - \alpha(\omega_{22} - \omega_{11}).$$

Time series homoskedasticity. If in addition to A1′ we assume that the marginal variance of v_{it} is constant for all periods:

$$E(v_{it}^2) = \sigma^2 \quad (t = 2, \ldots, T), \tag{A2}$$

it turns out that

$$\omega_{tt} = \alpha^2 \omega_{(t-1)(t-1)} + \phi + \sigma^2 + 2\alpha c_{t-1} \quad (t = 2, \ldots, T).$$

This gives rise to a covariance structure in which Ω depends on five free parameters: $\alpha, \phi, c_1, \omega_{11}, \sigma^2$. This is a model of some interest since it is one in which the initial

conditions of the process are unrestricted (governed by the parameters ϕ and c_1), but the total number of free parameters does not increase with T.

Mean stationarity of initial conditions. Other forms of additional structure that can be imposed are mean or variance stationarity conditions. The following assumption, which requires that the process started in the distant past, is a particularly useful mean stationarity condition:

$$\text{Cov}(y_{it} - y_{i(t-1)}, \eta_i) = 0 \quad (t = 2, \ldots, T). \tag{B1}$$

Relative to assumption A1, assumption B1 adds the following $(T-2)$ moment restrictions on Ω:

$$E[(y_{it}-\alpha y_{i(t-1)})\Delta y_{i(t-1)}]=0 \quad (t=3,\ldots,T), \tag{30}$$

which were proposed by Arellano and Bover (1995). However, relative to assumption A1′, assumption B1 only adds one moment restriction which can be written as $E[(y_{i3}-\alpha y_{i2})\Delta y_{i2}]=0$. In terms of the parameters c_t, the implication of assumption B1 is that $c_1=\cdots=c_T$ if we move from assumption A1, or that $c_1=\phi/(1-\alpha)$ if we move from assumption A1′. This gives rise to a model in which Ω depends on the $(T+2)\times 1$ parameter vector $\theta=(\alpha,\phi,\omega_{11},\ldots,\omega_{TT})'$. Notice that with $T=3$, α is overidentified under assumption B1. Now α will also satisfy

$$\alpha=(\omega_{22}-\omega_{21})^{-1}(\omega_{32}-\omega_{31}).$$

It is of some interest to note that the combination of assumptions A1 and B1 produces the same model as that of A1′ and B1. However, while A1′ implies orthogonality conditions that are quadratic in α, A1 or A1+B1 give rise to linear instrumental-variable conditions [Ahn and Schmidt (1995)]. While A1 implied the validity of lagged levels as instruments for equations in first-differences, B1 additionally implies the validity of lagged first-differences as instruments for equations in levels. The availability of instruments for levels equations may lead to the identification of the effect of observable components of η_i (i.e., time-invariant regressors), or to identifying unit roots, two points to which we shall return below.

The validity of assumption B1 depends on whether initial conditions at the start of the sample are representative of the steady state behaviour of the model or not. For example, for young workers or new firms initial conditions may be less related to steady state conditions than for older ones.

Full stationarity. By combining A1′ with the homoskedasticity and the mean stationarity assumptions, A2 and B1, we obtain a model whose only nonstationary feature is the variance of the initial observation, which would remain a free parameter. For such a model $\omega_{tt} = \alpha^2\omega_{(t-1)(t-1)} + \sigma^2 + \phi(1+\alpha)/(1-\alpha)$ $(t = 2, \ldots, T)$. A fully stationary specification results from making the additional assumption:

$$\omega_{11} = \frac{\phi}{(1-\alpha)^2} + \frac{\sigma^2}{(1-\alpha^2)}. \tag{B2}$$

This gives rise to a model in which Ω only depends on the three parameters α, ϕ, and σ^2. Nevertheless, identification still requires $T \geqslant 3$, despite the fact that with

$T = 2$, Ω has three different coefficients. To see this, note that in their relationship to α, ϕ, and σ^2 the equation for the second diagonal term is redundant:

$$\omega_{tt} = \sigma^2_{\eta*} + \sigma^2_{\ell} \quad (t = 1, 2), \quad \omega_{12} = \alpha(\omega_{11} - \sigma^2_{\eta*}) + \sigma^2_{\eta*},$$

where $\sigma^2_{\eta*} = \sigma^2_{\eta}/(1 - \alpha)^2$ and $\sigma^2_{\ell} = \sigma^2/(1 - \alpha^2)$. The intuition for this is that both η_i and $y_{i(t-1)}$ induce serial correlation on y_{it}, but their separate effects can only be distinguished if at least first and second order autocorrelations are observed.

Under full stationarity (assumptions A1, A2, B1, and B2) it can be shown that

$$\frac{E(\Delta y_{i\,(t+1)}\Delta y_{it})}{E[(\Delta y_{it})^2]} = -\frac{(1-\alpha)}{2}.$$

This is a well-known expression for the bias of the least squares regression in first-differences under homoskedasticity, which can be expressed as the orthogonality conditions

$$E\{\Delta y_{it}[(2y_{i\,(t+1)} - y_{it} - y_{i\,(t-1)}) - \alpha\Delta y_{it}]\} = 0 \quad (t = 2, \ldots, T-1).$$

With $T = 3$ this implies that α would also satisfy

$$\alpha = (\omega_{22} + \omega_{11} - 2\omega_{21})^{-1}[2(\omega_{32} - \omega_{31}) + \omega_{11} - \omega_{22}].$$

2.2.2. Aggregate shocks

Under assumptions A1 or A1′, the errors v_{it} are idiosyncratic shocks that are assumed to have cross-sectional zero mean at each point in time. However, if v_{it} contains aggregate shocks that are common to all individuals its cross-sectional mean will not be zero in general. This suggests replacing A1 with the assumption

$$E^*(v_{it} | y_i^{t-1}) = \delta_t \quad (t = 2, \ldots, T), \tag{31}$$

which leads to an extension of the basic specification in which an intercept is allowed to vary over time:

$$y_{it} = \delta_t + \alpha y_{i\,(t-1)} + \eta_i + \upsilon_{it}^{\dagger}, \tag{32}$$

where $\upsilon_{it}^{\dagger} = \upsilon_{it} - \delta_t$. We can now set $E(\eta_i) = 0$ without lack of generality, since a nonzero mean would be subsumed in δ_t. Again, formally Equation (32) is just a specialization of Equations (7) and (11).

With fixed T, this extension does not essentially alter the previous discussion since the realized values of the shocks δ_t can be treated as unknown period specific

parameters. With $T = 3$, α, δ_2 and δ_3 are just identified from the three moment conditions [10],

$$E(y_{i2} - \delta_2 - \alpha y_{i1}) = 0, \tag{33}$$

$$E(y_{i3} - \delta_3 - \alpha y_{i2}) = 0, \tag{34}$$

$$E[y_{i1}(\Delta y_{i3} - \Delta\delta_3 - \alpha\Delta y_{i2})] = 0. \tag{35}$$

In the presence of aggregate shocks the mean stationarity condition in assumption B1 may still be satisfied, but it will be interpreted as an assumption of mean stationarity conditional upon an aggregate effect (which may or may not be stationary), since now $E(\Delta y_{it})$ is not constant over t. The orthogonality conditions in Equation (30) remain valid in this case with the addition of a time varying intercept. With $T = 3$, assumption B1 adds to Equations (33–35) the orthogonality condition:

$$E[\Delta y_{i2}(y_{i3} - \delta_3 - \alpha y_{i2})] = 0. \tag{36}$$

2.2.3. *Identification and unit roots*

If one is interested in the unit root hypothesis, the model needs to be specified under both stable and unit roots environments. We begin by considering model (27) under assumption A1 as the stable root specification. As for the unit root specification, it is natural to consider a random walk without drift. The model can be written as

$$y_{it} = \alpha y_{i(t-1)} + (1-\alpha)\eta_i^* + v_{it}, \tag{37}$$

where η_i^* denotes the steady state mean of the process when $|\alpha| < 1$. Thus, when $\alpha = 1$ we have

$$y_{it} = y_{i(t-1)} + v_{it}, \tag{38}$$

so that heterogeneity only plays a role in the determination of the starting point of the process. Note that in this model the covariance matrix of (y_{i1}, η_i^*) is left unrestricted.

An alternative unit root specification would be a random walk with an individual specific drift given by η_i:

$$y_{it} = y_{i(t-1)} + \eta_i + \upsilon_{it}, \tag{39}$$

but this is a model with heterogeneous linear growth that would be more suited for comparisons with stationary models that include individual trends.

[10] Further discussion on models with time effects is contained in Crepon, Kramarz and Trognon (1997).

The main point to notice here is that in model (37) α is not identified from the moments derived from assumption A1 when $\alpha = 1$. This is so because in the unit root case the lagged level will be uncorrelated with the current innovation, so that $\mathrm{Cov}(y_{i(t-2)}, \Delta y_{i(t-1)}) = 0$. As a result, the rank condition will not be satisfied for the basic orthogonality conditions (28). In model (39) the rank condition is still satisfied since $\mathrm{Cov}(y_{i(t-2)}, \Delta y_{i(t-1)}) \neq 0$ due to the cross-sectional correlation induced by the heterogeneity in shifts.

As noted by Arellano and Bover (1995), this problem does not arise when we consider a stable root specification that in addition to assumption A1 satisfies the mean stationarity assumption B1. The reason is that when $\alpha = 1$ the moment conditions (30) remain valid and the rank condition is satisfied since $\mathrm{Cov}(\Delta y_{i(t-1)}, y_{i(t-1)}) \neq 0$.

2.2.4. *The value of information with highly persistent data*

The cross-sectional regression coefficient of y_{it} on $y_{i(t-1)}$, ρ_t, can be expressed as a function of the model's parameters. For example, under full stationarity it can be shown to be

$$\rho = \alpha + \frac{\mathrm{Cov}(\eta_i, y_{i(t-1)})}{\mathrm{Var}(y_{i(t-1)})} = \alpha + \frac{(1-\alpha)\lambda^2}{\lambda^2 + (1-\alpha)/(1+\alpha)} \geqslant \alpha \tag{40}$$

where $\lambda = \sigma_\eta/\sigma$. Often, empirically ρ is near unity. For example, with firm employment data, Alonso-Borrego and Arellano (1999) found $\rho = 0.995$, $\alpha = 0.8$, and $\lambda = 2$. Since for any $0 \leqslant \alpha \leqslant \rho$ there is a value of λ such that ρ equals a pre-specified value, in view of lack of identification of α from the basic moment conditions (28) when $\alpha = 1$, it is of interest to see how the information about α in these moment conditions changes as T and α change for values of ρ close to one.

For the orthogonality conditions (28) the inverse of the semiparametric information bound about α can be shown to be

$$\sigma_T^2 = \sigma^2 \left\{ \sum_{s=1}^{T-2} E(y_{is}^* y_i^{s\prime})[E(y_i^s y_i^{s\prime})]^{-1} E(y_i^s y_{is}^*) \right\}^{-1} \tag{41}$$

where the y_{is}^* are orthogonal deviations relative to $(y_{i1}, \ldots, y_{i(T-1)})'$ [11]. The expression σ_T^2 gives the lower bound on the asymptotic variance of any consistent estimator of α based exclusively on the moments (28) when the process generating the data is the fully stationary model [Chamberlain (1987)].

[11] That is, y_{is}^* is given by $y_{is}^* = c_s[y_{is} - (T-s-1)^{-1}(y_{i(s+1)} + \cdots + y_{i(T-1)})]$ $(s = 1, \ldots, T-2)$, where $c_s^2 = (T-s-1)/(T-s)$ [cf., Arellano and Bover (1995), and discussion in the next section].

Table 1
Inverse information bound for α (σ_T) when $\rho = 0.99$

T	σ_T [a]					
	(0, 9.9)	(0.2, 7.2)	(0.5, 4.0)	(0.8, 1.4)	(0.9, 0.7)	(0.99, 0)
3	14.14	15.50	17.32	18.97	19.49	19.95
4	1.97	2.66	4.45	8.14	9.50	10.00
5	1.21	1.55	2.43	4.71	5.88	6.34
10	0.50	0.57	0.71	1.18	1.61	1.85
15	0.35	0.38	0.44	0.61	0.82	0.96
Asympt. [b]	0.26	0.25	0.22	0.16	0.11	0.04

[a] Values for different (α, λ) pairs such that $\rho = 0.99$.
[b] Asymptotic standard deviation at $T = 15$, $\sqrt{(1-\alpha^2)/15}$.

In Table 1 we have calculated values of σ_T for various values of T and for different pairs (α, λ) such that $\rho = 0.99^{12}$. Also, the bottom row shows the time series asymptotic standard deviation, evaluated at $T = 15$, for comparisons.

Table 1 shows that with $\rho = 0.99$ there is a very large difference in information between $T = 3$ and $T > 3$. Moreover, for given T there is less information on α the closer α is to ρ. Often, there will be little information on α with $T = 3$ and the usual values of N. Additional information may be acquired from using some of the assumptions discussed above. Particularly, large gains can be obtained from employing mean stationarity assumptions, as suggested from Monte Carlo simulations reported by Arellano and Bover (1995) and Blundell and Bond (1998).

In making inferences about α we look for estimators whose sampling distribution for large N can be approximated by $N(\alpha, \sigma_T^2/N)$. However, there may be substantial differences in the quality of the approximation for a given N, among different estimators with the same asymptotic distribution. We shall return to these issues in the section on estimation.

2.3. *Using stationarity restrictions*

Some of the lessons from the previous section on alternative restrictions in autoregressive models are also applicable to regression models with predetermined (or strictly exogenous) variables of the form:

$$y_{it} = \delta' w_{it} + \eta_i + \upsilon_{it}, \tag{42}$$

$$E^*(\upsilon_{it} | w_i^t) = 0,$$

[12] Under stationarity σ_T^2 depends on α, λ and T but is invariant to σ^2.

where, e.g., $w_{it} = (y_{i(t-1)}, x_{it})'$. As before, the basic moments are $E[w_i^{t-1}(\Delta y_{it} - \delta' \Delta w_{it})] = 0$. However, if $E^*(v_{it}|w_i^t, \eta_i) = 0$ holds, the parameter vector δ also satisfies the Ahn–Schmidt restrictions

$$E[(y_{it} - \delta' w_{it})(\Delta y_{i(t-1)} - \delta \Delta w_{i(t-1)})] = 0. \tag{43}$$

Moreover, if $\mathrm{Cov}(\Delta w_{it}, \eta_i) = 0$ the Arellano–Bover restrictions are satisfied, encompassing the previous ones [13]:

$$E[\Delta w_{it}(y_{it} - \delta' w_{it})] = 0. \tag{44}$$

Blundell and Bond (1999) use moment restrictions of this type in their empirical analysis of Cobb–Douglas production functions using company panel data. They find that the instruments available for the production function in first differences are not very informative, due to the fact that the series on firm sales, capital and employment are highly persistent. In contrast, the first-difference instruments for production function errors in levels appear to be both valid and informative.

Sometimes the effect of time-invariant explanatory variables is of interest, a parameter γ, say, in a model of the form

$$y_{it} = \delta' w_{it} + \gamma z_i + \eta_i + \upsilon_{it}.$$

However, γ cannot be identified from the basic moments because the time-invariant regressor z_i is absorbed by the individual effect. Thus, we could ask whether the addition of orthogonality conditions involving errors in levels such as Equations (43) or (44) may help to identify such parameters. Unfortunately, often it would be difficult to argue that $E(\eta_i \Delta w_{it}) = 0$ without at the same time assuming that $E(z_i \Delta w_{it}) = 0$, in which case changes in w_{it} would not help the identification of γ. An example in which the levels restrictions may be helpful is the following simple model for an evaluation study due to Chamberlain (1993).

An evaluation of training example. Suppose that y_{it}^0 denotes earnings in the absence of training, and that there is a common effect of training for all workers. Actual earnings y_{it} are observed for $t = 1, \ldots, s-1, s+1, \ldots, T$. Training occurs in period s $(1 < s < T)$, so that $y_{it} = y_{it}^0$ for $t = 1, \ldots, s-1$, and we wish to measure its effect on earnings in subsequent periods, denoted by $\beta_{s+1}, \ldots, \beta_T$:

$$y_{it} = y_{it}^0 + \beta_t d_i \quad (t = s+1, \ldots, T), \tag{45}$$

where d_i is a dummy variable that equals 1 in the event of training. Moreover, we assume

$$y_{it}^0 = \alpha y_{i(t-1)}^0 + \eta_i + v_{it}, \tag{46}$$

[13] Strictly exogenous variables that had constant correlation with the individual effects were first considered by Bhargava and Sargan (1983).

together with $E^*(\upsilon_{it} | y_i^{0(t-1)}) = 0$ and $\mathrm{Cov}(\Delta y_{it}^0, \eta_i) = 0$. We also assume that d_i depends on lagged earnings $y_{i1}, \ldots, y_{i(s-1)}$ and η_i, but conditionally on these variables it is randomly assigned. Then we have:

$$y_{i(s+1)} = \alpha^2 y_{i(s-1)} + \beta_{s+1} d_i + (1+\alpha)\eta_i + (\upsilon_{i(s+1)} + \alpha \upsilon_{is}),$$

$$y_{it} = \alpha y_{i(t-1)} + (\beta_t - \alpha\beta_{t-1}) d_i + \eta_i + \upsilon_{it} \quad (t = s+2, \ldots, T).$$

From our previous discussion, the model implies the following orthogonality conditions:

$$E[y_i^{t-2}(\Delta y_{it} - \alpha \Delta y_{i(t-1)})] = 0 \quad (t = 1, \ldots, s-1), \tag{47}$$

$$E\{y_i^{s-2}[y_{i(s+1)} - (1+\alpha+\alpha^2)y_{i(s-1)} + \alpha(1+\alpha)y_{i(s-2)} - \beta_{s+1}d_i]\} = 0, \tag{48}$$

$$E\left\{y_i^{s-1}\left[y_{i(s+2)} - \frac{(1+\alpha+\alpha^2)}{(1+\alpha)}y_{i(s+1)} + \frac{\alpha^2}{(1+\alpha)}y_{i(s-1)} - \left(\beta_{i(s+2)} - \frac{(1+\alpha+\alpha^2)}{(1+\alpha)}\beta_{i(s+1)}\right)d_i\right]\right\} = 0. \tag{49}$$

$$E[y_i^{t-2}(\Delta y_{it} - \alpha \Delta y_{i(t-1)} + \Delta(\beta_t - \alpha\beta_{t-1})d_i)] = 0 \quad (t = s+3, \ldots, T). \tag{50}$$

The additional orthogonality conditions implied by mean stationarity are:

$$E[\Delta y_{i(t-1)}(y_{it} - \alpha y_{i(t-1)})] = 0 \quad (t = 1, \ldots, s-1), \tag{51}$$

$$E[\Delta y_{i(s-1)}(y_{i(s+1)} - \alpha^2 y_{i(s-1)} - \beta_{s+1} d_i)] = 0, \tag{52}$$

$$E[\Delta y_{i(s-1)}(y_{it} - \alpha y_{i(t-1)} + (\beta_t - \alpha\beta_{t-1}) d_i)] = 0 \quad (t = s+2, \ldots, T). \tag{53}$$

We would expect $E(\Delta y_{i(s-1)} d_i) < 0$, since there is evidence of a dip in the pretraining earnings of participants [e.g., Ashenfelter and Card (1985)]. Thus, Equation (52) can be expected to be more informative about β_{s+1} than Equation (48). Moreover, identification of β_{s+1} from Equation (48) requires that $s \geqslant 4$, otherwise only changes in β_t would be identified from Equations (47–50). In contrast, note that identification of β_{s+1} from Equation (52) only requires $s \geqslant 3$.

2.4. Models with multiplicative effects

In the models we have considered so far, unobserved heterogeneity enters exclusively through an additive individual specific intercept, while the other coefficients are assumed to be homogeneous. Nevertheless, an alternative autoregressive process could,

for example, specify a homogeneous intercept and heterogeneity in the autoregressive behaviour:

$$y_{it} = \gamma + (\alpha + \eta_i) y_{i(t-1)} + \upsilon_{it}.$$

This is a potentially useful model if one is interested in allowing for agent specific adjustment cost functions, as for example in labour demand models. If we assume $E(\upsilon_{it} | y_i^{t-1}) = 0$ and $y_{it} > 0$, the transformed model,

$$y_{it}\, y_{i(t-1)}^{-1} = \gamma\, y_{i(t-1)}^{-1} + \alpha + \eta_i + \upsilon_{it}^{+},$$

where $\upsilon_{it}^{+} = \upsilon_{it}\, y_{i(t-1)}^{-1}$, also has $E(\upsilon_{it}^{+} | y_i^{t-1}) = 0$. Thus, the average autoregressive coefficient α and the intercept γ can be determined in a way similar to the linear models from the moment conditions $E(\eta_i + \upsilon_{it}^{+}) = 0$ and $E(y_i^{t-2} \Delta \upsilon_{it}^{+}) = 0$. Note that in this case, due to the nonlinearity, the argument requires the use of conditional mean assumptions as opposed to linear projections.

Another example is an exponential regression of the form

$$E(y_{it}|x_i^t, y_i^{t-1}, \eta_i) = \exp(\beta x_{it} + \eta_i).$$

This case derives its motivation from the literature on Poisson models for count data. The exponential specification is chosen to ensure that the conditional mean is always non-negative. With count data a log-linear regression is not a feasible alternative since a fraction of the observations on y_{it} will be zeroes.

A third example is a model where individual effects are interacted with time effects given by

$$y_{it} = \beta x_{it} + \delta_t \eta_i + \upsilon_{it}.$$

A model of this type may arise in the specification of unrestricted linear projections as in Equations (21) and (22), or as a structural specification in which an aggregate shock δ_t is allowed to have individual-specific effects on y_{it} measured by η_i.

Clearly, in such multiplicative cases first-differencing does not eliminate the unobservable effects, but as in the heterogeneous autoregression above there are simple alternative transformations that can be used to construct orthogonality conditions.

A transformation for multiplicative models. Generalizing the previous specifications we have

$$f_t(w_i^T, \gamma) = g_t(w_i^t, \beta)\eta_i + \upsilon_{it}, \qquad E(\upsilon_{it}|w_i^t) = 0, \tag{54}$$

where $g_{it} = g_t(w_i^t, \beta)$ is a function of predetermined variables and unknown parameters such that $g_{it} > 0$ for all w_i^t and β, and $f_{it} = f_t(w_i^T, \gamma)$ depends on endogenous and

predetermined variables, as well as possibly also on unknown parameters. Dividing by g_{it} and first differencing the resulting equation, we obtain

$$f_{i(t-1)} - (g_{it}^{-1} g_{i(t-1)}) f_{it} = v_{it}^{+}, \tag{55}$$

and

$$E(v_{it}^{+} | w_i^{t-1}) = 0.$$

where $v_{it}^{+} = v_{i(t-1)} - (g_{it}^{-1} g_{i(t-1)}) v_{it}$.

Any function of w_i^{t-1} will be uncorrelated with v_{it}^+ and therefore can be used as an instrument in the determination of the parameters β and γ. This kind of transformation has been suggested by Chamberlain (1992b) and Wooldridge (1997). Notice that its use does not require us to condition on η_i. However, it does require g_t to be a function of predetermined variables as opposed to endogenous variables.

Multiple individual effects. We turn to consider models with more than one heterogeneous coefficient. Multiplicative random effects models with strictly exogenous variables were considered by Chamberlain (1992a), who found the information bound for a model with a multivariate individual effect. Chamberlain (1993) considered the identification problems that arise in models with predetermined variables when the individual effect is a vector with two or more components, and showed lack of identification of α in a model of the form

$$y_{it} = \alpha y_{i(t-1)} + \beta_i x_{it} + \eta_i + v_{it}, \tag{56}$$

$$E(v_{it} | x_i^t, y_i^{t-1}) = 0 \quad (t = 2, \ldots, T). \tag{57}$$

As an illustration consider the case where x_{it} is a 0 – 1 binary variable. Since $E(\eta_i|x_i^T, y_i^{T-1})$ is unrestricted, the only moments that are relevant for the identification of α are

$$E(\Delta y_{it} - \alpha \Delta y_{i(t-1)}|x_i^{t-1}, y_i^{t-2}) = E(\beta_i \Delta x_{it}|x_i^{t-1}, y_i^{t-2}) \quad (t = 3, \ldots, T).$$

Letting $w_i^t = (x_i^t, y_i^t)$, the previous expression is equivalent to the following two conditions:

$$\begin{aligned} E(\Delta y_{it} - \alpha \Delta y_{i(t-1)}|w_i^{t-2}, x_{i(t-1)} = 0) = {} & E(\beta_i|w_i^{t-2}, x_{i(t-1)} = 0) \\ & \times \Pr(x_{it} = 1|w_i^{t-2}, x_{i(t-1)} = 0), \end{aligned} \tag{58}$$

$$\begin{aligned} E(\Delta y_{it} - \alpha \Delta y_{i(t-1)}|w_i^{t-2}, x_{i(t-1)} = 1) = {} & -E(\beta_i|w_i^{t-2}, x_{i(t-1)} = 1) \\ & \times \Pr(x_{it} = 0|w_i^{t-2}, x_{i(t-1)} = 1). \end{aligned} \tag{59}$$

Clearly, if $E(\beta_i|w_i^{t-2}, x_{i(t-1)} = 0)$ and $E(\beta_i|w_i^{t-2}, x_{i(t-1)} = 1)$ are unrestricted, and T is fixed, the autoregressive parameter α cannot be identified from Equations (58) and (59).

Let us consider some departures from model (56–57) under which α would be potentially identifiable. Firstly, if x were a strictly exogenous variable, in the sense that we replaced Equation (57) with the assumption $E(\upsilon_{it}|x_i^T, y_i^{t-1}) = 0$, α could be identifiable since

$$E(\Delta y_{it} - \alpha \Delta y_{i(t-1)}|x_i^T, y_i^{t-2}, \Delta x_{it} = 0) = 0. \tag{60}$$

Secondly, if the intercept η were homogeneous, identification of α and η could result from

$$E(y_{it} - \eta - \alpha y_{i(t-1)}|w_i^{t-1}, x_{it} = 0) = 0. \tag{61}$$

The previous discussion illustrates the fragility of the identification of dynamic responses from short time series of heterogeneous cross-sectional populations.

If $x_{it} > 0$ in model (56–57), it may be useful to discuss the ability of transformation (55) to produce orthogonality conditions. In this regard, a crucial aspect of the previous case is that while x_{it} is predetermined in the equation in levels, it becomes endogenous in the equation in first differences, so that transformation (55) applied to the first-difference equation does not lead to conditional moment restrictions. The problem is that although $E(\Delta v_{it} | x_i^{t-1}, y_i^{t-2}) = 0$, in general $E[(\Delta x_{it})^{-1} \Delta v_{it} | x_i^{t-1}, y_i^{t-2}] \neq 0$.

The parameters α, $\beta = E(\beta_i)$, and $\gamma = E(\eta_i)$ could be identifiable if x were a strictly exogenous variable such that $E(\upsilon_{it}|x_i^T, y_i^{t-1}) = 0$ $(t = 2, \ldots, T)$, for in this case the transformed error $\upsilon_{it}^+ = (\Delta x_{it})^{-1}\Delta\upsilon_{it}$ would satisfy $E[\upsilon_{it}^+|x_i^T, y_i^{t-2}] = 0$ and $E[\Delta\upsilon_{it}^+|x_i^T, y_i^{t-3}] = 0$. Therefore, the following moment conditions would hold:

$$E\left[\left(\frac{\Delta y_{it}}{\Delta x_{it}} - \frac{\Delta y_{i(t-1)}}{\Delta x_{i(t-1)}}\right) - \alpha\left(\frac{\Delta y_{i(t-1)}}{\Delta x_{it}} - \frac{\Delta y_{i(t-2)}}{\Delta x_{i(t-1)}}\right) \,\middle|\, x_i^T, y_i^{t-3}\right] = 0, \tag{62}$$

$$E\left(\frac{\Delta y_{it}}{\Delta x_{it}} - \alpha\frac{\Delta y_{i(t-1)}}{\Delta x_{it}} - \beta\right) = 0, \tag{63}$$

$$\begin{aligned} E\Bigg[&\left(\frac{\Delta(y_{it}/x_{it})}{\Delta(1/x_{it})} - \frac{\Delta(y_{i(t-1)}/x_{i(t-1)})}{\Delta(1/x_{i(t-1)})}\right) \\ &- \alpha\left(\frac{\Delta(y_{i(t-1)}/x_{it})}{\Delta(1/x_{it})} - \frac{\Delta(y_{i(t-2)}/x_{i(t-1)})}{\Delta(1/x_{i(t-1)})}\right) \,\Bigg|\, x_i^T, y_i^{t-3}\Bigg] = 0, \end{aligned} \tag{64}$$

$$E\left(\frac{\Delta(y_{it}/x_{it})}{\Delta(1/x_{it})} - \alpha\frac{\Delta(y_{i(t-1)}/x_{it})}{\Delta(1/x_{it})} - \gamma\right) = 0. \tag{65}$$

A similar result would be satisfied if x_{it} in Equation (56) were replaced by a predetermined regressor that remained predetermined in the equation in first differences like $x_{i(t-1)}$. The result is that transformation (55) could be sequentially applied to models with predetermined variables and multiple individual effects, and still produce orthogonality conditions, as long as T is sufficiently large, and the

transformed model resulting from the last but one application of the transformation still has the general form (54) (i.e., no functions of endogenous variables are multiplied by individual specific parameters).

A heterogeneous AR(1) model. As another example, consider a heterogeneous AR(1) model for a 0 – 1 binary indicator y_{it}:

$$y_{it} = \eta_i + \alpha_i y_{i(t-1)} + \upsilon_{it}, \tag{66}$$
$$E(\upsilon_{it} | y_i^{t-1}) = 0,$$

and let us examine the (lack of) identification of the expected autoregressive parameter $E(\alpha_i)$ and the expected intercept $E(\eta_i)$. With $T = 3$, the only moment that is relevant for the identification of $E(\alpha_i)$ is

$$E(\Delta y_{i3} | y_{i1}) = E(\alpha_i \Delta y_{i2} | y_{i1}),$$

which is equivalent to the following two conditions:

$$E(\Delta y_{i3}|y_{i1}=0)=E(\alpha_i|y_{i1}=0,y_{i2}=1)\Pr(y_{i2}=1|y_{i1}=0), \tag{67}$$

$$E(\Delta y_{i3}|y_{i1}=1)=-E(\alpha_i|y_{i1}=1,y_{i2}=0)\Pr(y_{i2}=0|y_{i1}=1). \tag{68}$$

Therefore, only $E(\alpha_i|y_{i1}=0,y_{i2}=1)$ and $E(\alpha_i|y_{i1}=1,y_{i2}=0)$ are identified. The expected value of α_i for those whose value of y does not change from period 1 to period 2 is not identified, and hence $E(\alpha_i)$ is not identified either.

Similarly, for $T > 3$ we have

$$\begin{aligned} E(\Delta y_{it} | y_i^{t-3}, y_{i(t-2)} = 0) &= E(\alpha_i | y_i^{t-3}, y_{i(t-2)} = 0,\ y_{i(t-1)} = 1) \\ &\quad \times \Pr(y_{i(t-1)} = 1 | y_i^{t-3}, y_{i(t-2)} = 0), \\ E(\Delta y_{it} | y_i^{t-3}, y_{i(t-2)} = 1) &= -E(\alpha_i | y_i^{t-3}, y_{i(t-2)} = 1,\ y_{i(t-1)} = 0) \\ &\quad \times \Pr(y_{i(t-1)} = 0 | y_i^{t-3}, y_{i(t-2)} = 1). \end{aligned}$$

Note that $E(\alpha_i | y_i^{t-3}, y_{i(t-2)} = j,\ y_{i(t-1)} = j)$ for $j = 0, 1$ is also identified provided $E(\alpha_i | y_i^{t-3}, y_{i(t-2)} = j)$ is identified on the basis of the first $T - 1$ observations. The conclusion is that all conditional expectations of α_i are identified except $E(\alpha_i | y_{i1} = \cdots = y_{i(T-1)} = 1)$ and $E(\alpha_i | y_{i1} = \cdots = y_{i(T-1)} = 0)$.

Concerning η_i, note that since $E(\eta_i|y_i^{T-1}) = E(y_i^T|y_i^{T-1}) - y_{i(T-1)}E(\alpha_i|y_i^{T-1})$, expectations of the form $E(\eta_i|y_i^{T-2}, y_{i(T-1)} = 0)$ are all identified. Moreover, $E(\eta_i|y_i^{T-2}, y_{i(T-1)} = 1)$ is identified provided $E(\alpha_i|y_i^{T-2}, y_{i(T-1)} = 1)$ is identified. Thus, all conditional expectations of η_i are identified except $E(\eta_i|y_{i1} = \cdots = y_{i(T-1)} = 1)$.

Note that if $\Pr(y_{i1} = \cdots = y_{i(T-1)} = j)$ for $j = 0, 1$ tends to zero as T increases, $E(\alpha_i)$ and $E(\eta_i)$ will be identified as $T \to \infty$, but they may be seriously underidentified for very small values of T.

3. Linear models with predetermined variables: estimation

3.1. GMM estimation

Consider a model for panel data with sequential moment restrictions given by

$$\begin{aligned} y_{it} &= x_{it}'\beta_o + u_{it} \quad (t = 1, \ldots, T;\ i = 1, \ldots, N), \\ u_{it} &= \eta_i + \upsilon_{it}, \qquad E^*(\upsilon_{it} \mid z_i^t) = 0 \end{aligned} \tag{69}$$

where x_{it} is a $k \times 1$ vector of possibly endogenous variables, z_{it} is a $p \times 1$ vector of instrumental variables, which may include current values of x_{it} and lagged values of y_{it} and x_{it}, and $z_i^t = (z_{i1}', \ldots, z_{it}')'$. Observations across individuals are assumed to be independent and identically distributed. Alternatively, we can write the system of T equations for individual i as

$$y_i = X_i\beta_o + u_i, \tag{70}$$

where $y_i = (y_{i1}, \ldots, y_{iT})'$, $X_i = (x_{i1}', \ldots, x_{iT}')'$, and $u_i = (u_{i1}, \ldots, u_{iT})'$.

We saw that this model implies instrumental-variable orthogonality restrictions for the model in first-differences. In fact, the restrictions can be expressed using any $(T-1)\times T$ upper-triangular transformation matrix K of rank $(T-1)$, such that $K\iota = 0$, where ι is a $T \times 1$ vector of ones. Note that the first-difference operator is an example. We then have

$$E(Z_i' K u_i) = 0, \tag{71}$$

where Z_i is a block-diagonal matrix whose tth block is given by $z_i^{t\prime}$. An optimal GMM estimator of β_o based on Equation (71) is given by

$$\widehat{\beta} = (M_{zx}' A M_{zx})^{-1} M_{zx}' A M_{zy}, \tag{72}$$

where $M_{zx} = \left(\sum_{i=1}^{N} Z_i' K X_i\right)$, $M_{zy} = \left(\sum_{i=1}^{N} Z_i' K y_i\right)$, and A is a consistent estimate of the inverse of $E(Z_i' K u_i u_i' K' Z_i)$ up to a scalar. Under "classical" errors (that is, under conditional homoskedasticity $E(v_{it}^2 \mid z_i^t) = \sigma^2$, and lack of autocorrelation $E(v_{it}\, v_{i(t+j)} \mid z_i^{t+j}) = 0$ for $j > 0$), a "one-step" choice of A is optimal:

$$A_C = \left(\sum_{i=1}^{N} Z_i' K K' Z_i\right)^{-1}. \tag{73}$$

Alternatively, the standard "two-step" robust choice is

$$A_{\mathrm{R}} = \left(\sum_{i=1}^{N} Z_i' K \tilde{u}_i \tilde{u}_i' K' Z_i \right)^{-1}, \tag{74}$$

where $\tilde{u}_i = y_i - X_i \tilde{\beta}$ is a vector of residuals evaluated at some preliminary consistent estimate $\tilde{\beta}$.

Given identification, $\widehat{\beta}$ is consistent and asymptotically normal as $N \to \infty$ for fixed T [Hansen (1982)]. In addition, for either choice of A, provided the conditions under which they are optimal choices are satisfied, the asymptotic variance of $\widehat{\beta}$ is

$$\mathrm{Var}(\widehat{\beta})_{\mathrm{R}} = \{E(X_i'K'Z_i)[E(Z_i'K\,u_i\,u_i'K'Z_i)]^{-1}E(Z_i'K\,X_i)\}^{-1}, \tag{75}$$

which is invariant to K. Under classical errors this becomes [14]

$$\mathrm{Var}(\widehat{\beta})_{\mathrm{C}} = \sigma^2\{E(X_i'K'Z_i)[E(Z_i'K\,K'Z_i)]^{-1}E(Z_i'K\,X_i)\}^{-1}.$$

Moreover, as shown by Arellano and Bover (1995), a GMM estimator of the form given in Equations (72) and (73) or (74), is invariant to the choice of K provided K satisfies the required conditions [see also Schmidt, Ahn and Wyhowski (1992)].

As in common with other GMM estimation problems, the minimized estimation criterion provides an asymptotic chi-squared test statistic of the overidentifying restrictions. A two-step Sargan test statistic is given by

$$S_{\mathrm{R}} = \left[\sum_{i=1}^{N}(y_i - X_i\widehat{\beta}_{\mathrm{R}})'K'Z_i\right] A_{\mathrm{R}} \left[\sum_{i=1}^{N} Z_i'K(y_i - X_i\widehat{\beta}_{\mathrm{R}})\right] \rightarrow \chi^2_{(q-k)}, \tag{76}$$

where $\widehat{\beta}_{\mathrm{R}}$ is the two-step GMM estimator [15].

Orthogonal deviations. An alternative transformation to first differencing, which is very useful in the context of models with predetermined variables, is forward orthogonal deviations:

$$u_{it}^{*} = c_t \left[u_{it} - \frac{1}{(T-t)} (u_{i\,(t+1)} + \cdots u_{iT}) \right], \tag{77}$$

where $c_t^2 = (T-t)/(T-t+1)$ [Arellano and Bover (1995)]. That is, to each of the first $(T-1)$ observations we subtract the mean of the remaining future observations available in the sample. The weighting c_t is introduced to equalize the variances of the transformed errors. A closely related transformation was used by Hayashi and Sims (1983) for time series models.

Unlike first differencing, which introduces a moving average structure in the error term, orthogonal deviations preserve lack of correlation among the transformed errors if the original ones are not autocorrelated and have constant variance. Indeed,

[14] Under classical errors, additional moment restrictions would be available, with the result that a smaller asymptotic variance could be achieved. The expression above simply particularizes the asymptotic variance to a situation where additional properties occur in the population but are not used in estimation.

[15] Similarly, letting $\widehat{\sigma}^2$ and $\widehat{\beta}_C$ be, respectively, a consistent estimate of σ^2 and the one-step estimator, the one-step Sargan statistic is given by $S_C = \widehat{\sigma}^{-2} \left[\sum_{i=1}^{N} (y_i - X_i\widehat{\beta}_C)'K'Z_i\right] A_C \left[\sum_{i=1}^{N} Z_i'K(y_i - X_i\widehat{\beta}_C)\right]$.

orthogonal deviations can be regarded as the result of doing first differences to eliminate fixed effects plus a GLS transformation to remove the serial correlation induced by differencing.

The choice of K that produces this transformation is the forward orthogonal deviations operator $A = \text{diag}[(T-1)/T, \ldots, 1/2]^{1/2} A^{+}$, where

$$A^{+} = \begin{pmatrix} 1 & -(T-1)^{-1} & -(T-1)^{-1} & \cdots & -(T-1)^{-1} & -(T-1)^{-1} & -(T-1)^{-1} \\ 0 & 1 & -(T-2)^{-1} & \cdots & -(T-2)^{-1} & -(T-2)^{-1} & -(T-2)^{-1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1/2 & -1/2 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

It can be verified by direct multiplication that $AA' = I_{(T-1)}$ and $A'A = I_T - \iota\iota'/T \equiv Q$, which is the within-group operator. Thus, the OLS regression of y_{it}^* on x_{it}^* will give the within-group estimator, which is the conventional estimator in static models with strictly exogenous variables. Finally, since $Q = K'(KK')^{-1}K$, also $A = (KK')^{-1/2}K$ for any upper-triangular K.

A useful computational feature of orthogonal deviations, specially so when T is not a very small number, is that one-step estimators can be obtained as a matrix-weighted average of cross-sectional IV estimators:

$$\widehat{\beta} = \left(\sum_{t=1}^{T-1} X_t^{*\prime} Z_t (Z_t'Z_t)^{-1} Z_t' X_t^* \right)^{-1} \sum_{t=1}^{T-1} X_t^{*\prime} Z_t (Z_t'Z_t)^{-1} Z_t' y_t^*, \tag{78}$$

where $X_t^* = (x_{1t}^{*\prime}, \ldots, x_{Nt}^{*\prime})'$, $y_t^* = (y_{1t}^*, \ldots, y_{Nt}^*)'$, and $Z_t = (z_i^{t\prime}, \ldots, z_N^{t\prime})'$.

An illustration: female labour force participation and fertility. We illustrate the previous issues with reference to an empirical relationship between female participation and fertility, discussing a simplified version of the results reported by Carrasco (1998) for a linear probability model [16].

A sample from PSID for 1986–1989 is used. The data consists of 1442 women aged 18–55 in 1986, that are either married or cohabiting. The left-hand side variable is a binary indicator of participation in year t. Fertility is also a dummy variable, which takes the value one if the age of the youngest child in $t + 1$ is 1. The equation also includes an indicator of whether the woman has a child aged 2–6. The equations estimated in levels also include a constant, age, race, and education dummies (not reported).

In this sample it is observed that women with two children of the same sex have a significantly higher probability of having a third child. Thus, the sex of the first two children is used as an instrument for fertility, which is treated as an endogenous

[16] We thank Raquel Carrasco for allowing us to draw freely on her dataset and models.

Table 2
Linear probability models of female labour force participation[a,b] (N = 1442, 1986–1989)

Variable	OLS	2SLS[c]	WITHIN	GMM[d]	GMM[e]
Fertility	–0.15	–1.01	–0.06	–0.08	–0.13
	(8.2)	(2.1)	(3.8)	(2.8)	(2.2)
Kids 2–6	–0.08	–0.24	0.001	–0.005	–0.09
	(5.2)	(2.6)	(0.04)	(0.4)	(2.7)
Sargan test				48.0 (22)	18.0 (10)
*m*1	19.0	5.7	–10.0	–10.0	–10.0
*m*2	16.0	12.0	–1.7	–1.7	–1.6
Models including lagged participation					
Fertility	–0.09	–0.33	–0.06	–0.09	–0.14
	(5.2)	(1.3)	(3.7)	(3.1)	(2.2)
Kids 2–6	–0.02	–0.07	–0.000	–0.02	–0.10
	(2.1)	(1.3)	(0.00)	(1.1)	(3.5)
Lagged participation	0.63	0.61	0.03	0.36	0.29
	(42.0)	(30.0)	(1.7)	(8.3)	(6.3)
Sargan				51.0 (27)	25.0 (15)
*m*1	–7.0	–5.4	–13.0	–14.0	–13.0
*m*2	3.1	2.8	–1.3	1.5	1.2

[a] Heteroskedasticity robust t-ratios shown in parentheses.
[b] GMM IVs in bottom panel also include lags of participation up to $t-2$.
[c] External instrument: previous children of same sex.
[d] IVs: all lags and leads of "kids 2–6" and "same sex" variables (strictly exogenous).
[e] IVs: lags of "kids 2–6" and "same sex" up to $t-1$ (predetermined).

variable. The presence of a child aged 2–6 is the result of past fertility decisions, and so it should be treated as a predetermined variable [see Carrasco (1998) for a comprehensive discussion, and additional estimates of linear and nonlinear models].

Table 2 reports the results for two versions of the model with and without lagged participation as a regressor, using DPD [Arellano and Bond (1988)]. The last column presents GMM estimates in orthogonal deviations that treat fertility as endogenous, and the "kids 2–6" and "same sex" indicators as predetermined variables. The table also reports the results from other methods of estimation for comparisons.

There is a large gap between the OLS and 2SLS measured effects of fertility, possibly due to measurement errors. Both OLS and 2SLS neglect unobserved heterogeneity, despite evidence from the serial correlation statistics $m1$ and $m2$ of persistent positive autocorrelation in the residuals in levels. Note that we would expect the "same sex" instrumental variable to be correlated with the fixed effect. The reason

is that it will be a predictor of preferences for children, given that the sample includes women with less than two children.

The within-groups estimator controls for unobserved heterogeneity, but in doing so we would expect it to introduce biases due to lack of strict exogeneity of the explanatory variables. The GMM estimates in column 4 deal with the endogeneity of fertility and control for fixed effects, but treat the "kids 2–6" and "same sex" variables as strictly exogenous. This results in a smaller effect of fertility on participation (in absolute value) than the one obtained in column 5 treating the variables as predetermined. The hypothesis of strict exogeneity of these two variables is rejected at the 5 percent level from the difference in the Sargan statistics in both panels. (Both GMM estimates are "one-step", but all test statistics reported are robust to heteroskedasticity.)

Finally, note that the $m1$ and $m2$ statistics (which are asymptotically distributed as a $N(0, 1)$ under the null of no autocorrelation) have been calculated from residuals in first differences for the within-groups and GMM estimates. So if the errors in levels were uncorrelated, we would expect $m1$ to be significant, but not $m2$, as is the case here [cf., Arellano and Bond (1991)].

Levels and differences estimators. The GMM estimator proposed by Arellano and Bover (1995) combined the basic moments (71) with $E(\Delta z_{it} u_{it}) = 0$ $(t = 2, \ldots, T)$. Using their notation, the full set of orthogonality conditions can be written in compact form as

$$E(Z_i^{+\prime} H u_i) = 0, \tag{79}$$

where Z_i^+ is a block diagonal matrix with blocks Z_i as above, and $Z_{\ell i} = \text{diag}\,(\Delta z_{i2}', \ldots, \Delta z_{iT}')$. H is the $2(T-1) \times T$ selection matrix $H = (K', I_o')'$, where $I_o = (0 : I_{T-1})$. With these changes in notation, the form of the estimator is similar to that in Equation (72).

As before, a robust choice of A is provided by the inverse of an unrestricted estimate of the variance matrix of the moments $N^{-1}\sum_{i=1}^{N} Z_i^{+\prime} H \tilde{u}_i \tilde{u}_i' H' Z_i^{+}$. However, this can be a poor estimate of the population moments if N is not sufficiently large relative to T, which may have an adverse effect on the finite sample properties of the GMM estimator. Unfortunately, in this case an efficient one-step estimator under restrictive assumptions does not exist. Intuitively, since some of the instruments for the equations in levels are not valid for those in differences, and conversely, not all the covariance terms between the two sets of moments will be zero.

3.2. Efficient estimation under conditional mean independence

If lack of correlation between v_{it} and z_i^t is replaced by an assumption of conditional independence in mean $E(v_{it} \mid z_i^t) = 0$, the model implies additional orthogonality restrictions. This is so because v_{it} will be uncorrelated not only with the conditioning

variables z_i^t but also with functions of them. Chamberlain (1992b) derived the semiparametric efficiency bound for this model. Hahn (1997) showed that a GMM estimator based on an increasing set of instruments as N tends to infinity would achieve the semiparametric efficiency bound. Hahn discussed the rate of growth of the number of instruments for the case of Fourier series and polynomial series.

Note that the asymptotic bound for the model based on $E(v_{it} \mid z_i^t) = 0$ will be in general different from that of $E(v_{it} \mid z_i^t, \eta_i) = 0$, whose implications for linear projections were discussed in the previous section.

Similarly, the bound for a version of the model with levels and differences restrictions based on conditional mean independence assumptions cannot be obtained either as an application of Chamberlain's results. The reason is that the addition of the level's conditions breaks the sequential moment structure of the problem.

Let us now consider the form of the information bound and the optimal instruments for model (69) together with the conditional mean assumption $E(v_{it} \mid z_i^t) = 0$. Since $E(\eta_i \mid z_i^T)$ is unrestricted, all the information about β is contained in $E(v_{it} - v_{i(t+1)} \mid z_i^t) = 0$ for $t = 1, \ldots, T-1$.

For a single period the information bound is $J_{0t} = E(d_{it}\, d_{it}'/\omega_{it})$ where $d_{it} = E(x_{it} - x_{i(t+1)} \,|\, z_i^t)$ and $\omega_{it} = E[(v_{it} - v_{i(t+1)})^2 \,|\, z_i^t]$ [cf., Chamberlain (1987)]. Thus, for a single period the optimal instrument is $m_{it} = d_{it}/\omega_{it}$, in the sense that under suitable regularity conditions the statistic

$$\tilde{\beta}_{(t)} = \left(\sum_{i=1}^{N} m_{it}\, \Delta x_{i(t+1)}' \right)^{-1} \left(\sum_{i=1}^{N} m_{it}\, \Delta y_{i(t+1)} \right),$$

satisfies $\sqrt{N}(\widetilde{\beta}_{(t)} - \beta) \xrightarrow{d} N(0, J_{0t}^{-1})$. If the errors were conditionally serially uncorrelated, the total information would be the sum of the information bounds for each period. So Chamberlain (1992b) proposed the following recursive forward transformation of the first-differenced errors:

$$
\begin{aligned}
\tilde{v}_{i(T-1)} &= v_{i(T-1)} - v_{iT}, \\
\tilde{v}_{it} &= (v_{it} - v_{i(t+1)}) \\
&\quad - \frac{E[(v_{it} - v_{i(t+1)})\tilde{v}_{i(t+1)} | z_i^{t+1}]}{E(\tilde{v}_{i(t+1)}^2 | z_i^{t+1})} \tilde{v}_{i(t+1)} \\
&\quad - \frac{E[(v_{it} - v_{i(t+1)})\tilde{v}_{i(t+2)} | z_i^{t+2}]}{E(\tilde{v}_{i(t+2)}^2 | z_i^{t+2})} \tilde{v}_{i(t+2)} \\
&\quad - \cdots \\
&\quad - \frac{E[(v_{it} - v_{i(t+1)})\tilde{v}_{i(T-1)} | z_i^{T-1}]}{E(\tilde{v}_{i(T-1)}^2 | z_i^{T-1})} \tilde{v}_{i(T-1)},
\end{aligned}
\tag{80}
$$

for $t = T-2, \ldots, 1$. The interest in this transformation is that it satisfies the same conditional moment restrictions as the original errors in first-differences, namely

$$
E(\tilde{v}_{it} | z_i^t) = 0, \tag{81}
$$

but additionally it satisfies by construction the lack of dependence requirement:

$$E(\tilde{v}_{it}\,\tilde{v}_{i\,(t+j)}\,|\,z_i^{t+j}) = 0 \text{ for } j = 1, \ldots, T-t-1. \tag{82}$$

Therefore, in terms of the transformed errors the information bound can be written as

$$J_0 = \sum_{t=1}^{T-1} E(\tilde{d}_{it}\,\tilde{d}'_{it}/\tilde{\omega}_{it}), \tag{83}$$

where $\widetilde{d}_{it} = E(\tilde{x}_{it} \,|\, z_i^t)$ and $\widetilde{\omega}_{it} = E(\tilde{v}_{it}^2 \,|\, z_i^t)$. The variables $\tilde{x}_{it}$ and $\tilde{y}_{it}$ denote the corresponding transformations to the first-differences of x_{it} and y_{it} such that $\tilde{v}_{it} = \tilde{y}_{it} - \tilde{x}_{it}'\beta$. Thus, the optimal instruments for all periods are $\widetilde{m}_{it} = \widetilde{d}_{it}/\widetilde{\omega}_{it}$, in the sense that under suitable regularity conditions the statistic

$$\widetilde{\beta} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T-1} \widetilde{m}_{it} \tilde{x}_{it}' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T-1} \widetilde{m}_{it} \tilde{y}_{it} \right)$$

satisfies $\sqrt{N}(\widetilde{\beta} - \beta) \xrightarrow{d} N(0, J_0^{-1})$.

If the υ_{it}'s are conditionally homoskedastic and serially uncorrelated, so that $E(\upsilon_{it}^2 \mid z_i^t) = \sigma^2$ and $E(\upsilon_{it}\upsilon_{i(t+j)} \mid z_i^{t+j}) = 0$ for $j > 0$, it can be easily verified that the $\tilde{\upsilon}_{it}$'s blow down to ordinary forward orthogonal deviations as defined in Equation (77):

$$\tilde{\upsilon}_{it} = \upsilon_{it} - \frac{1}{(T-t)}(\upsilon_{i(t+1)} + \cdots + \upsilon_{iT}) \equiv \frac{1}{c_t}\upsilon_{it}^* \text{ for } t = T-1, \ldots, 1.$$

In such case $\tilde{m}_{it} = c_t\sigma^{-2}E(x_{it}^* \mid z_i^t)$ so that

$$\tilde{\beta} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T-1} E(x_{it}^* \mid z_i^t)x_{it}^{*\prime}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T-1} E(x_{it}^* \mid z_i^t)y_{it}^*\right), \tag{84}$$

and

$$J_0 = \frac{1}{\sigma^2}\sum_{t=1}^{T-1} E[E(x_{it}^* \mid z_i^t)E(x_{it}^{*\prime} \mid z_i^t)]. \tag{85}$$

If we further assume that the conditional expectations $E(x_{it}^{*} \mid z_i^t)$ are linear, then

$$J_0 = \frac{1}{\sigma^2} \sum_{t=1}^{T-1} E(x_{it}^{*} z_i^{t\prime})[E(z_i^t z_i^{t\prime})]^{-1} E(z_i^t x_{it}^{*\prime}), \tag{86}$$

which coincides with the inverse of the asymptotic covariance matrix of the simple IV estimator given in Equation (78) under the stated assumptions. Note that the

assumptions of conditional homoskedasticity, lack of serial correlation, and linearity of $E(x_{it}^* \mid z_i^t)$ would imply further conditional moment restrictions that may lower the information bound for β. Here, we merely particularize the bound for β based on $E(v_{it} \mid z_i^t) = 0$ to the case where the additional restrictions happen to occur in the population but are not used in the calculation of the bound.

3.3. *Finite sample properties of GMM and alternative estimators*

For sufficiently large N, the sampling distribution of the GMM estimators discussed above can be approximated by a normal distribution. However, the quality of the approximation for a given sample size may vary greatly depending on the quality of the instruments used. Since the number of instruments increases with T, many overidentifying restrictions tend to be available even for moderate values of T, although the quality of these instruments is often poor.

Monte Carlo results on the finite sample properties of GMM estimators for panel data models with predetermined variables have been reported by Arellano and Bond (1991), Kiviet (1995), Ziliak (1997), Blundell and Bond (1998) and Alonso-Borrego and Arellano (1999), amongst others. A conclusion in common to these studies is that GMM estimators that use the full set of moments available for errors in first-differences can be severely biased, specially when the instruments are weak and the number of moments is large relative to the cross-sectional sample size.

From the literature on the finite sample properties of simultaneous equations estimators, we know that the effect of weak instruments on the distributions of 2SLS and LIML differs substantially, in spite of the fact that both estimators have the same asymptotic distribution. While LIML is approximately median unbiased, 2SLS is biased towards OLS, and in the case of lack of identification in the population it converges to a random variable with the OLS probability limit as its central value. In contrast, LIML has no moments, and as a result its distribution has thicker tails than that of 2SLS and a higher probability of outliers [cf., Phillips (1983)]. Anderson, Kunitomo and Sawa (1982) carried out numerical comparisons of the distributions of the two estimators, and concluded that LIML was to be strongly preferred to 2SLS, specially in cases with a large number of instruments.

LIML analogue estimators. It is thus of interest to consider LIML analogues for our models, and compare their finite sample properties with those of GMM estimators. Following Alonso-Borrego and Arellano (1999), a non-robust LIML analogue $\widehat{\beta}_{\mathrm{LIML1}}$ minimizes a criterion of the form

$$\ell_{\mathrm{C}}(\beta) = \frac{(y^* - X^*\beta)'M(y^* - X^*\beta)}{(y^* - X^*\beta)'(y^* - X^*\beta)}, \tag{87}$$

where starred variables denote orthogonal deviations, $y^* = (y_1^{*\prime}, \ldots, y_N^{*\prime})'$, $X^* = (X_1^{*\prime}, \ldots, X_N^{*\prime})'$, $Z = (Z_1', \ldots, Z_N')'$, and $M = Z(Z'Z)^{-1}Z'$. The resulting estimator is

$$\widehat{\beta}_{\mathrm{LIML1}} = (X^{*\prime}MX^* - \hat{\ell}X^{*\prime}X^*)^{-1}(X^{*\prime}My^* - \hat{\ell}X^{*\prime}y^*), \tag{88}$$

where $\hat{\ell}$ is the minimum eigenvalue of the matrix $W^{*\prime}M\,W^{*}(W^{*\prime}W^{*})^{-1}$, and $W^{*} = (y^{*},X^{*})$.

The estimator in Equation (88) is algebraically similar to an ordinary single-equation LIML estimator provided the model is in orthogonal deviations. This is so in spite of having a system of equations, due to the fact that the errors in orthogonal deviations of different equations are serially uncorrelated and homoskedastic under classical assumptions. However, the non-robust LIML analogue does not correspond to any meaningful maximum likelihood estimator (for example, it does not exploit the homoskedasticity restrictions). It is only a "LIML" estimator in the sense of the instrumental-variable interpretation given by Sargan (1958) to the original LIML estimator, and generalized to robust contexts by Hansen, Heaton and Yaron (1996).

The robust LIML analogue $\widehat{\beta}_{\mathrm{LIML2}}$, or continuously updated GMM estimator in the terminology of Hansen et al. (1996), minimizes a criterion of the form

$$\ell_{\mathrm{R}}(\beta) = (y^* - X^*\beta)'Z\left(\sum_{i=1}^{N} Z_i' u_i^*(\beta) u_i^*(\beta)' Z_i\right)^{-1} Z'(y^* - X^*\beta), \tag{89}$$

where $u_i^*(\beta) = y_i^* - X_i^*\beta$. Note that LIML2, unlike LIML1, does not solve a standard minimum eigenvalue problem, and requires the use of numerical optimization methods [17].

In contrast to GMM, the LIML estimators are invariant to normalization. Hillier (1990) showed that the alternative normalization rules adopted by LIML and 2SLS were at the root of their different sampling properties. He also showed that a symmetrically normalized 2SLS estimator had similar properties to those of LIML. Alonso-Borrego and Arellano (1999) considered symmetrically normalized GMM (SNM) estimators for panel data, and compared them with ordinary GMM and LIML analogues by mean of simulations. The main advantage of robust SNM over robust LIML is computational, since the former solves a minimum eigenvalue problem while the latter does not. It also avoids potential problems of non-convergence with LIML2, as reported by Alonso-Borrego and Arellano (1999).

The Monte Carlo results and the empirical illustrations for autoregressive models reported by Alonso-Borrego and Arellano (1999) showed that GMM estimates can exhibit large biases when the instruments are poor, while the symmetrically normalized estimators (LIML and SNM) remained essentially unbiased. However, LIML and SNM always had a larger interquartile range than GMM, although the differences were small except in the almost unidentified cases.

[17] Other one-step methods that achieve the same asymptotic efficiency as robust GMM or LIML estimators are the empirical likelihood [Back and Brown (1993), Qin and Lawless (1994) and Imbens (1997)] and exponential tilting estimators [Imbens, Spady and Johnson (1998)]. Nevertheless, little is known as yet on the relative merits of these estimators in panel data models, concerning computational aspects and their finite sample properties.

3.4. *Approximating the distributions of GMM and LIML for AR(1) models when the number of moments is large*

Within-groups estimators of autoregressive models, and more generally of models with predetermined variables, are known to be consistent as T tends to infinity, but are inconsistent for fixed T and large N [cf., Nickell (1981), Anderson and Hsiao (1981)]. On the other hand, the estimators reviewed above are consistent for fixed T but the number of orthogonality conditions increases with T. In panels in which the value of T is not negligible relative to N (such as the PSID household incomes panel in the US, or the balance sheet-based company panels that are available in many countries), the knowledge of the asymptotic behaviour of the estimators as both T and N tend to infinity may be useful in assessing alternative methods.

Alvarez and Arellano (1998) obtained the asymptotic properties of within-groups (WG), one-step GMM, and non-robust LIML for a first-order autoregressive model when both N and T tend to infinity. Hahn (1998) also obtained the asymptotic properties of WG under more general conditions. The main results can be summarized in the following proposition.

Proposition 1. *Let* $y_{it} = \alpha y_{i(t-1)} + \eta_i + \upsilon_{it}$, *with* $\upsilon_{it}|y_i^{t-1}, \eta_i \sim$ *i.i.d.*$N(0, \sigma^2)$, $(t = 1, \ldots, T)$ *and* $y_{i0}|\eta_i \sim N[\eta_i/(1-\alpha), \sigma^2/(1-\alpha^2)]$. *Also let* $\eta_i \sim$ *i.i.d.*$N(0, \sigma_\eta^2)$. *Then, as both N and T tend to infinity, provided* $T/N \to c$, $0 \leqslant c \leqslant 2$, *within-groups, GMM1, and LIML1 are consistent for* α. *Moreover,*

$$\sqrt{NT}\left[\widehat{\alpha}_{\mathrm{GMM1}} - \left(\alpha - \frac{1}{N}(1+\alpha)\right)\right] \xrightarrow{d} N(0, 1-\alpha^2), \tag{90}$$

$$\sqrt{NT}\left[\widehat{\alpha}_{\mathrm{LIML1}} - \left(\alpha - \frac{1}{(2N-T)}(1+\alpha)\right)\right] \xrightarrow{d} N(0, 1-\alpha^2). \tag{91}$$

Also, provided $N/T^3 \to 0$:

$$\sqrt{NT}\left[\widehat{\alpha}_{\mathrm{WG}} - \left(\alpha - \frac{1}{T}(1+\alpha)\right)\right] \xrightarrow{d} N(0, 1-\alpha^2). \tag{92}$$

Proof: See Alvarez and Arellano (1998) [18].

The consistency result contrasts with those available for the structural equation setting, where 2SLS is inconsistent when the ratio of number of instruments to sample size tends to a positive constant [cf., Kunitomo (1980), Morimune (1983), Bekker (1994)]. Here the number of instruments, which is given by $T(T-1)/2$, increases very fast and yet consistency is obtained. The intuition for this result is that in our context as

[18] Here, for notational convenience, we assume that y_{i0} is also observed, so that the effective number of time series observations will be $T+1$.

T tends to infinity the "simultaneity bias" tends to zero, and so closeness of GMM1 or LIML1 to OLS in orthogonal deviations (ie. within-groups) becomes a desirable property.

Note that when $T/N \rightarrow 0$ the fixed T results for GMM1 and LIML1 remain valid, but within-groups, although consistent, has an asymptotic bias in its asymptotic distribution (which would only disappear if $N/T \rightarrow 0$). However, when T/N tends to a positive constant, within-groups, GMM1 and LIML1 exhibit negative biases in their asymptotic distributions. The condition that $c > 2$ is not restrictive since GMM1 and LIML1 are only well defined for $(T-1)/N \leqslant 1$. Thus, for $T < N$ the GMM1 bias is always smaller than the within-groups bias, and the LIML1 bias is smaller than the other two.

Another interesting feature is that the three estimators are asymptotically efficient in the sense of attaining the same asymptotic variance as the within-groups estimator as $T \to \infty$. However, Alvarez and Arellano (1998) show that the standard formulae for fixed T estimated variances of GMM1 and LIML1, which depend on the variance of the fixed effect, remain consistent estimates of the asymptotic variances as $T \to \infty$.

These results provide some theoretical support for LIML1 over GMM1. They also illustrate the usefulness of understanding the properties of panel data estimators as the time series information accumulates, even for moderate values of T: in a fixed T framework, GMM1 and LIML1 are asymptotically equivalent, but as T increases LIML1 has a smaller asymptotic bias than GMM1.

The crude GMM estimator in first differences. Alvarez and Arellano (1998) also show that the crude GMM estimator (CIV) that neglects the autocorrelation in the first differenced errors (ie., one-step GMM in first-differences with weight matrix equal to $(Z'Z)^{-1}$) is inconsistent as $T/N \to c > 0$, despite being consistent for fixed T. The result is:

$$\widehat{\alpha}_{\mathrm{CIV}} \xrightarrow{p} \alpha - \frac{(1+\alpha)}{2}\left(\frac{c}{2-(1+\alpha)(2-c)/2}\right). \tag{93}$$

The intuition for this result is that the "simultaneity bias" of OLS in first differences (unlike the one for orthogonal deviations) does not tend to zero as $T \to \infty$. Thus, for fixed T the IV estimators in orthogonal deviations and first differences are both consistent, whereas as T increases the former remains consistent but the latter is inconsistent. Moreover, notice that the bias may be qualitatively relevant. Standard fixed-T large-N GMM theory would just describe the CIV estimator as being asymptotically less efficient than GMM1 as a consequence of using a non-optimal choice of weighting matrix.