Panel Data Models: Some Recent Developments

Manuel Arellano
CEMFI

Bo Honoré
Princeton University

Working Paper No. 0016
November 2000

We thank Jason Abrevaya, Badi Baltagi, Olympia Bover, Martin Browning, Jim Heckman, Luojia Hu, Ekaterini Kyriazidou, Ed Leamer, Enrique Sentana, Jeffrey Wooldridge and participants at the Chicago and London Handbook conferences for helpful comments. All errors are our responsibility. (Email addresses: arellano@cemfi.es, honore@princeton.edu).

CEMFI, Casado del Alisal 5, 28014 Madrid, Spain.
www.cemfi.es

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.

JEL Classification: C33

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:

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

We assume that $(y_{i1}...y_{iT}, x_{i1}...x_{iT}, \eta_i)$ 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 v_{it}:

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

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

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

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 $T \geq 2$.

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

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

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

$$\begin{aligned} y_{i1} &= \pi_{01} + \beta x_{i1} + \lambda' x_i^T + \varepsilon_{i1} \quad E^*(\varepsilon_{i1}|x_i^T) = 0 & (5) \\ \Delta y_{it} &= \Delta\delta_t + \beta \Delta x_{it} + \Delta\varepsilon_{it} \qquad E^*(\Delta\varepsilon_{it}|x_i^T) = 0 \ (t = 2, ..., T). & (6) \end{aligned}$$

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 (4) is equivalent to asking under which conditions β is identified in (6).[1]

[1] Lack of dependence between v_{it} and x_i^T could also be expressed in terms of conditional independence in mean $E(v_{it}|x_i^T) = 0$ $(t = 1, ..., T)$. In the absence of any knowledge about η_i this is equivalent to the $(T - 1)$ conditional moment restrictions $E(v_{it} - v_{i(t-1)}|x_i^T) = 0$ $(t = 2, ..., T)$ 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:

$$y_{it} = \alpha y_{i(t-1)} + \beta_0 x_{it} + \beta_1 x_{i(t-1)} + \delta_t + \eta_i + v_{it} \; (i = 1, ..., N; t = 2, ..., T) \tag{7}$$

together with

$$E^*(v_{it}|x_i^T) = 0 \; (t = 2, ..., T). \tag{8}$$

Note that assumption (8) does not restrict the serial correlation of v, 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 v_{it} since they may be autocorrelated in an unspecified way. Likewise, the system in first differences is free from fixed effects and satifies $E^*(\Delta v_{it}|x_i^T) = 0$ $(t = 3, ..., T)$, but $\Delta y_{i(t-1)}$ may still be correlated with Δv_{is} for all s.

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

$$\begin{aligned} E[\begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \\ x_{i3} \end{pmatrix} (\Delta y_{i3} - \alpha \Delta y_{i2} - \beta_0 \Delta x_{i3} - \beta_1 \Delta x_{i2} - \Delta \delta_3)] &= 0 \\ E\left(y_{i2} - \alpha y_{i1} - \beta_0 x_{i2} - \beta_1 x_{i1} - \delta_2\right) &= 0. \end{aligned} \tag{9}$$

This simple example illustrates the potential for cross-sectional identification of 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 (eg. 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 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:

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

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

[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.

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 (7) if

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

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

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

Examples of this situation include Euler equations for household consumption (Zeldes, 1989, Runkle, 1991, and Keane and Runkle, 1992), or for company investment (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 (11) for errors in first differences is that

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

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 $T \geq 3$. With $T = 3$ they are just identified from the five orthogonality conditions:

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

It is of some interest to compare the situation in (13) with that in (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 (8) were replaced with:

$$E^{*}(v_{it}|x_{i}^{T},y_{i}^{t-1})=0\ (t=2,...,T). \tag{14}$$

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 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

[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.

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, 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 (11) we have

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

In this case serial correlation is not ruled out, and the partial adjustment model is identifiable with $T \geq 4$, but (15) rules out unspecified feedback from lagged y to current x. As an example, suppose that $v_{it} = \zeta_{it} + \varepsilon_{it}$ 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 v's will be serially correlated and correlated with lagged consumption variables y but not with lagged price variables x. Another example is an equation $y_{it}^* = \beta x_{it} + \eta_i + v_{it}^*$ where v_{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.

[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).

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

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

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 (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 case. The moment conditions for the partial adjustment model with a predetermined x uncorrelated with the effects can be written as

$$\begin{aligned} E[\begin{pmatrix} 1 \\ x_i^t \end{pmatrix} \left(y_{it} - \alpha y_{i(t-1)} - \beta_0 x_{it} - \beta_1 x_{i(t-1)} - \delta_t\right)] &= 0 \ (t = 2, ..., T) \quad (18) \\ E[y_i^{t-2} \left(\Delta y_{it} - \alpha \Delta y_{i(t-1)} - \beta_0 \Delta x_{it} - \beta_1 \Delta x_{i(t-1)} - \Delta \delta_t\right)] &= 0 \ (t = 3, ..., T) \end{aligned}$$

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

[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).

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 $x_i^{(t+1)T}=(x_{i(t+1)},...,x_{iT})'$ 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 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 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

[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).

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

$$\begin{aligned} w_{it} &= w_{it}^{p} + \varepsilon_{it} && (23) \\ c_{it} &= \beta w_{it}^{p} + \xi_{it} && (24) \\ x_{it} &= \gamma w_{it}^{p} + \varsigma_{it}. && (25) \end{aligned}$$

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

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

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

[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.

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

[8] In general, linear conditional models can be represented as data covariance matrix structures, but

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.

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

Let us consider a random sample of individual time series of size T, $\{y_i^T, i = 1, ..., N\}$, 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:

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

$$E^*(v_{it} | y_i^{t-1}) = 0 \; (t = 2, ..., T) \tag{A1}$$

where $E(\eta_i) = \gamma$, $E(v_{it}^2) = \sigma_t^2$, and $Var(\eta_i) = \sigma_\eta^2$. Notice that the assumption does not rule out correlation between η_i and v_{it}, nor the possibility of conditional heteroskedasticity, since $E(v_{it}^2 | y_i^{t-1})$ need not coincide with σ_t^2. (27) and (A1) can be seen as a specialization of (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 (27) by y_{is} for $s < t$, and taking expectations gives $\omega_{ts} = \alpha\omega_{(t-1)s} + c_s$, $(t = 2, ..., T; s = 1, ..., t-1)$, 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 $\theta = (\alpha, c_1, ..., c_{T-1}, \omega_{11}, ..., \omega_{TT})'$. Notice that with $T = 3$ the parameters (α, c_1, c_2) are just identified as functions of the elements of Ω:

$$\alpha = (\omega_{21} - \omega_{11})^{-1}(\omega_{31} - \omega_{21})$$

typically they involve a larger parameter space including many nuisance parameters, which are absent from instrumental-variable orthogonality conditions.

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

$$\begin{aligned} c_1 &= \omega_{21} - \alpha\omega_{11} \\ c_2 &= \omega_{32} - \alpha\omega_{22}. \end{aligned}$$

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 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

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

Note that in a short panel assumption A1$'$ is more restrictive than assumption A1. Nevertheless, lack of correlation between v_{it} and $\{y_{i(t-1)}, ..., y_{i(t-J)}\}$ implies lack of correlation between v_{it} and η_i in the limit as $J \to \infty$. This will be so as long as

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

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

When $T \geq 4$, assumption A1$'$ implies the following additional $T-3$ quadratic moment restrictions that were considered by Ahn and Schmidt (1995):

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

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 $E(\eta_i \Delta v_{i(t-1)}) = 0$ the result follows. Thus, (29) also holds if $Cov(\eta_i, v_{it})$ is constant over t.

An alternative representation of the restrictions in (29) is in terms of a recursion of the coefficients c_t introduced above. Multiplying (27) by η_i and taking expectations gives $c_t = \alpha c_{t-1} + \phi, (t = 2, ..., T)$, where $\phi = E(\eta_i^2) = \gamma^2 + \sigma_\eta^2$, so that $c_1, ..., c_T$ 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 $\theta = (\alpha, \phi, c_1, \omega_{11}, ..., \omega_{TT})'$. 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:

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

it turns out that

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

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:

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

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

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

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 $c_1 = ... = c_T$ 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 $\theta = (\alpha, \phi, \omega_{11}, ..., \omega_{TT})'$. 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 (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 model $\omega_{tt} = \alpha^2\omega_{(t-1)(t-1)} + \sigma^2 + \phi(1+\alpha)/(1-\alpha)$ $(t = 2, ..., T)$. 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 $T \geq 3$, 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:

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

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

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

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

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

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

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

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

$$\begin{aligned} E(y_{i2} - \delta_2 - \alpha y_{i1}) &= 0 & (33) \\ E(y_{i3} - \delta_3 - \alpha y_{i2}) &= 0 & (34) \\ E[y_{i1}(\Delta y_{i3} - \Delta\delta_3 - \alpha\Delta y_{i2})] &= 0. & (35) \end{aligned}$$

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 (30) remain valid in this case with the addition of a time varying intercept. With $T = 3$, (B1) adds to (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

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

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

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:

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

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

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 $Cov(y_{i(t-2)}, \Delta y_{i(t-1)}) = 0$. 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 $Cov(y_{i(t-2)}, \Delta y_{i(t-1)}) \neq 0$ 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 (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 $Cov(\Delta y_{i(t-1)}, y_{i(t-1)}) \neq 0$.

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

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

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 $0 \leq \alpha \leq \rho$ 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 $(y_{i1}, ..., y_{i(T-1)})'$.[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).

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
Inverse Information Bound for α (σ_T)
$\rho = 0.99$

α	0	0.2	0.5	0.8	0.9	0.99
λ	9.9	7.2	4.0	1.4	0.7	0
$T = 3$	14.14	15.50	17.32	18.97	19.49	19.95
$T = 4$	1.97	2.66	4.45	8.14	9.50	10.00
$T = 5$	1.21	1.55	2.43	4.71	5.88	6.34
$T = 10$	0.50	0.57	0.71	1.18	1.61	1.85
$T = 15$	0.35	0.38	0.44	0.61	0.82	0.96
$\left(\frac{1-\alpha^2}{15}\right)^{1/2}$	0.26	0.25	0.22	0.16	0.11	0.04

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

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

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

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:

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

$$E^*(v_{it}|w_i^t) = 0$$

where, for example, $w_{it} = (y_{i(t-1)}, x_{it})'$. 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 $Cov(\Delta w_{it}, \eta_i) = 0$ 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

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

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

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

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 (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 $t = 1, ..., s-1, s+1, ..., T$. Training occurs in period s $(1 < s < T)$, so that $y_{it} = y_{it}^0$ for $t = 1, ..., s-1$, and we wish to measure its effect on earnings in subsequent periods, denoted by $\beta_{s+1}, ..., \beta_T$:

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

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

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

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

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

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

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

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

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

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

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

The additional orthogonality conditions implied by mean stationarity are:

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

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

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

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 (eg. Ashenfelter and Card, 1985). Thus, (52) can be expected to be more informative about β_{s+1} than (48). Moreover, identification of β_{s+1} from (48) requires that $s \geq 4$, otherwise only changes in β_t would be identified from (47)-(50). In contrast, note that identification of β_{s+1} from (52) only requires $s \geq 3$.

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:

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

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 $E(v_{it}|y_i^{t-1}) = 0$ and $y_{it} > 0$, the transformed model

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

where $v_{it}^+ = v_{it}y_{i(t-1)}^{-1}$, also has $E(v_{it}^+|y_i^{t-1}) = 0$. 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 $E(\eta_i + v_{it}^+) = 0$ and $E(y_i^{t-2}\Delta v_{it}^+) = 0$. 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

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

A model of this type may arise in the specification of unrestricted linear projections as in (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

$$f_t(w_i^T, \gamma) = g_t(w_i^t, \beta)\eta_i + v_{it} \tag{54}$$

$$E(v_{it}|w_i^t) = 0$$

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

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

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

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

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

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

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

and

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

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 (57) with the assumption $E(v_{it} | x_i^T, y_i^{t-1}) = 0$, 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 $E(v_{it} | x_i^T, y_i^{t-1}) = 0$ $(t = 2, ..., T)$, for in this case the transformed error $v_{it}^{+} = (\Delta x_{it})^{-1}\Delta v_{it}$ would satisfy $E[v_{it}^{+}|x_i^T, y_i^{t-2}] = 0$ and $E[\Delta v_{it}^{+}|x_i^T, y_i^{t-3}] = 0$. Therefore, the following moment conditions would hold:

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

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

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

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

A similar result would be satisfied if x_{it} in (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}:

$$y_{it} = \eta_i + \alpha_i y_{i(t-1)} + v_{it} \tag{66}$$

$$E(v_{it}|y_i^{t-1}) = 0,$$

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:

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

and

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

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

$$E(\Delta y_{it}|y_i^{t-3}, y_{i(t-2)}=0) = E(\alpha_i|y_i^{t-3}, y_{i(t-2)}=0, y_{i(t-1)}=1)\Pr(y_{i(t-1)}=1|y_i^{t-3}, y_{i(t-2)}=0)$$

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

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 $E(\alpha_i|y_{i1}=...=y_{i(T-1)}=1)$ and $E(\alpha_i|y_{i1}=...=y_{i(T-1)}=0)$.

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 $E(\eta_i|y_{i1}=...=y_{i(T-1)}=1)$.

Note that if $\Pr(y_{i1}=...=y_{i(T-1)}=j)$ 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

$$y_{it} = x_{it}'\beta_o + u_{it}\ (t=1,...,T; i=1,...,N) \tag{69}$$

$$u_{it} = \eta_i + v_{it}$$

$$E^*(v_{it}|z_i^t) = 0$$

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 $z_i^t = (z_{i1}', ..., z_{it}')'$. 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 $y_i = (y_{i1}, ..., y_{iT})'$, $X_i = (x_{i1}', ..., x_{iT}')'$, and $u_i = (u_{i1}, ..., u_{iT})'$.

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 t-th block is given by $z_i^{t\prime}$. An optimal GMM estimator of β_o based on (71) is given by

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

where $M_{zx} = (\sum_{i=1}^N Z_i' K X_i)$, $M_{zy} = (\sum_{i=1}^N Z_i' K y_i)$, 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 $E(v_{it}^2|z_i^t) = \sigma^2$, and lack of autocorrelation $E(v_{it} v_{i(t+j)}|z_i^{t+j}) = 0$ 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

$$A_R = \left(\sum_{i=1}^N Z_i' K \widetilde{u}_i \widetilde{u}_i' K' Z_i \right)^{-1} \tag{74}$$

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

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

$$Var(\widehat{\beta})_R = \{E(X_i'K'Z_i)[E(Z_i'Ku_iu_i'K'Z_i)]^{-1}E(Z_i'KX_i)\}^{-1}, \tag{75}$$

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

$$Var(\widehat{\beta})_C = \sigma^2\{E(X_i'K'Z_i)[E(Z_i'KK'Z_i)]^{-1}E(Z_i'KX_i)\}^{-1}.$$

Moreover, as shown by Arellano and Bover (1995), a GMM estimator of the form given in (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

$$S_R = \left[\sum_{i=1}^{N}(y_i - X_i\widehat{\beta}_R)'K'Z_i\right] A_R \left[\sum_{i=1}^{N} Z_i'K(y_i - X_i\widehat{\beta}_R)\right] \to \chi^2_{(q-k)} \tag{76}$$

where $\widehat{\beta}_R$ 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:

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

[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 $S_C = \widehat{\sigma}^{-2}\left[\sum_{i=1}^{N}(y_i - X_i\widehat{\beta}_C)'K'Z_i\right] A_C \left[\sum_{i=1}^{N} Z_i'K(y_i - X_i\widehat{\beta}_C)\right]$

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, 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 $A = diag[(T-1)/T, ..., 1/2]^{1/2}A^{+}$, 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:

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

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

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 1986-89 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 variable. The presence of a child 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

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

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.)

Table 2
Female Labour Force Participation
Linear Probability Models (N=1442, 1986-1989)

Variable	OLS	2SLS[1]	WITHIN	GMM[2]	GMM[3]
				(St.Exog.)	(Predet.)
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. (22)	18. (10)
$m1$	19.	5.7	-10.	-10.	-10.
$m2$	16.	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 Partic.	0.63	0.61	0.03	0.36	0.29
	(42.)	(30.)	(1.7)	(8.3)	(6.3)
Sargan				51. (27)	25. (15)
$m1$	-7.0	-5.4	-13.	-14.	-13.
$m2$	3.1	2.8	-1.3	1.5	1.2

Heteroskedasticity robust t-ratios shown in parentheses.
[1] External instrument: Previous children of same sex.
[2] IVs: All lags and leads of "kids 2-6" and "same sex" variables.
[3] IVs: Lags of "kids 2-6" and "same sex" up to $t-1$.
GMM IVs in bottom panel also include lags of partic. up to $t-2$.

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 & Differences Estimators The GMM estimator proposed by Arellano and Bover (1995) combined the basic moments (71) with $E(\Delta z_{it} u_{it}) = 0$, $(t = 2, ..., T)$. 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 $Z_{\ell i} = diag(\Delta z_{i2}', ..., \Delta z_{iT}')$. H is the $2(T-1) \times T$ selection matrix $H = (K', I_o')'$, where $I_o = (0 \vdots I_{T-1})$. With this changes in notation, the form of the estimator is similar to that in (72).

As before, a robust choice of A is provided by the inverse of an unrestricted estimate of the variance matrix of the moments $N^{-1} \sum_{i=1}^{N} Z_i^{+\prime} H \widetilde{u}_i \widetilde{u}_i' H' Z_i^+$. 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 $E(v_{it}|z_i^t) = 0$, 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 $E(v_{it}|z_i^t) = 0$ will be in general different from that of $E(v_{it}|z_i^t, \eta_i) = 0$, 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 $E(v_{it}|z_i^t) = 0$. Since $E(\eta_i|z_i^T)$ is unrestricted, all the information about β is contained in $E(v_{it} - v_{i(t+1)}|z_i^t) = 0$ for $t = 1, ..., T-1$.

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

$$\widetilde{v}_{i(T-1)} = v_{i(T-1)} - v_{iT}$$

$$\begin{aligned}\widetilde{v}_{it} &= (v_{it} - v_{i(t+1)}) - \frac{E[(v_{it} - v_{i(t+1)})\widetilde{v}_{i(t+1)}|z_i^{t+1}]}{E(\widetilde{v}_{i(t+1)}^2|z_i^{t+1})}\widetilde{v}_{i(t+1)} - \frac{E[(v_{it} - v_{i(t+1)})\widetilde{v}_{i(t+2)}|z_i^{t+2}]}{E(\widetilde{v}_{i(t+2)}^2|z_i^{t+2})}\widetilde{v}_{i(t+2)} \\ &\quad - ... - \frac{E[(v_{it} - v_{i(t+1)})\widetilde{v}_{i(T-1)}|z_i^{T-1}]}{E(\widetilde{v}_{i(T-1)}^2|z_i^{T-1})}\widetilde{v}_{i(T-1)} \qquad (80)\end{aligned}$$

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

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

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

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

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

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

where $\widetilde{d}_{it} = E(\widetilde{x}_{it}|z_i^t)$ and $\widetilde{\omega}_{it} = E(\widetilde{v}_{it}^2|z_i^t)$. The variables $\widetilde{x}_{it}$ and $\widetilde{y}_{it}$ denote the corresponding transformations to the first-differences of x_{it} and y_{it} such that $\widetilde{v}_{it} = \widetilde{y}_{it} - \widetilde{x}'_{it}\beta$. 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 $\widetilde{\beta} = \left(\sum_{i=1}^N \sum_{t=1}^{T-1} \widetilde{m}_{it}\widetilde{x}'_{it}\right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^{T-1} \widetilde{m}_{it}\widetilde{y}_{it}\right)$ satisfies $\sqrt{N}(\widetilde{\beta} - \beta) \xrightarrow{d} N(0, J_0^{-1})$.

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

$$\widetilde{v}_{it} = v_{it} - \frac{1}{(T-t)}(v_{i(t+1)} + ... + v_{iT}) \equiv \frac{1}{c_t}v_{it}^* \text{ for } t = T-1, ..., 1.$$

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

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

and

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

If we further assume that the conditional expectations $E(x_{it}^*|z_i^t)$ 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 (78) under the stated assumptions. Note that the assumptions of conditional homoskedasticity, lack of serial correlation, and linearity of $E(x_{it}^*|z_i^t)$ would imply further conditional moment restrictions that may lower the information bound for β. Here, we merely particularize the bound for β based on $E(v_{it}|z_i^t) = 0$ 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, 1998, 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 $\widehat{\beta}_{LIML1}$ minimizes a criterion of the form

$$\ell_C(\beta) = \frac{(y^* - X^*\beta)' M (y^* - X^*\beta)}{(y^* - X^*\beta)'(y^* - X^*\beta)} \tag{87}$$

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

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

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

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

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

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]

[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.

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 (1998) 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.

The Monte Carlo results and the empirical illustrations for autoregressive models reported by Alonso-Borrego and Arellano 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.

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

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

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

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

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

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, and 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 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 $T/N \to 0$ the fixed T results for GMM1 and LIML1 remain valid, but within-groups, although consistent, has an asymptotic bias in its asymptotic distribution

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

(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 $(T-1)/N \leq 1$. 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 $T \rightarrow \infty$. However, Alvarez and Arellano 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 $T \rightarrow \infty$.

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 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 $T/N \rightarrow c > 0$, despite being consistent for fixed T. The result is:

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

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 $T \rightarrow \infty$. 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.

4 Nonlinear Panel Data Models

The ability to difference out the individual specific effect as was done in the previous sections relies heavily on the linear or multiplicative way in which it entered the model. Many simple cross sectional models have a constant that does not enter in this way. This is for example true for all the limited dependent variable models discussed in Chapters 9 and 10 of Amemiya (1985). Introducing an individual specific effect as an individual specific constant in those models, therefore results in models that cannot be estimated by the methods discussed so far. As will be seen in the following sections, the currently available methods for dealing with these models, rely on insights that are model–specific and that do not always seem to be useful for similar, but slightly different models. The main exception to this is the conditional maximum likelihood approach which has been used to construct estimators for some exponential family models. We discuss this method in the next section.

Unfortunately, there are many models for which it is not possible to use the conditional likelihood approach to eliminate the individual specific effect. For some of those models, alternative appoaches have been developed. In sections 6 and 7, we will review some of the progress that has been made in the area of estimation of limited dependent variable models with individual–specific, "fixed", effects[19]. This literature is closely related to the literature that deals with estimation of semiparametric limited dependent variables models, in that it is usually not necessary to specify a parametric form for the distribution of the underlying errors. The models are also semiparametric in the sense that the distribution of the individual specific effects conditional on the explanatory variable, is left unspecified. It is therefore not surprising that there is a close relationship between some of the approaches that are discussed here, and some approaches that have been taken to estimation of semiparametric limited dependent variables mod-

[19] Even though one often imagines a random sample of individuals, and hence random draws of the individual specifc effects, it is customary to call the effect "fixed" when no assumptions are made on its relationship with other explanatory variables. A random effect is one which has been modelled in some manner.

els. Indeed, in some cases the estimators for the panel data models have preceded the "corresponding" estimators for the cross sectional models.

The main limitation of much of the literature on nonlinear panel data methods, is that it is assumed that the explanatory variables are strictly exogenous in the sense that some assumptions will be made on the errors conditional on all (including future) values of the explanatory variables. As was pointed out earlier in this chapter, many of the recent advances in estimation of linear panel data models have focused on relaxing this assumption. In section 8, we will discuss how some of the methods can be generalized to allow for lagged dependent variables, but at this point very little is known about estimation of nonlinear panel data models with predetermined explanatory variables.

The discussion of nonlinear panel data models in the next three sections will focus entirely on standard nonlinear econometric models in which the parameter that is usually interpreted as an intercept, is allowed to be individual specific. This seems like a natural first step in understanding the value and limitations of panel data when the model of interest is nonlinear. However, it is clear that knowing the "parameters of interest" in the models discussed below does not always allow one to infer all the quantities of interest. For example, in the fixed effects logit model below, knowing β will not allow one to infer the effect of one of the explanatory variables on the probability distribution of the dependent variable, although knowing the vector of β's will allow one to infer the relative effects of the explanatory variables. This problem is due to the semiparametric nature of the nonlinear models considered here, and is not particular to panel data. On the other hand, if the censoring in equation (103) below is due to top– or bottom– coding of the true dependent variable of interest, then the interpretation of the parameters of the censored regression model is exactly the same as the interpretation of the parameters of a linear panel data model. The same can sometimes be said for the selection models discussed below.

Another limitation of most of the discussion here is that it focuses on the extreme case where no assumptions are made on the relationship between the individual specific effect and the explanatory variables. Whether a more "random" effects approach where some assumptions are made on how the distribution of this effect depends on the explanatory variables is more useful, depends on the context (and one's taste). In section 9 we briefly discuss some recent advances in this area. We devote much less space to that topic because many of the new developments there are by–products of developments in other areas of econometrics. For example, recent developments in Bayesian econometrics and in simulation–based inference have implications for nonlinear random effects panel data models, but the main new insights are more general, and not really tied to panel data.

5 Conditional Maximum Likelihood Estimation

In a static linear model, one can justify treating the individual specific effects as parameters to be estimated by reference to the Frisch–Waugh Theorem: OLS (or normal maximum likelihood) on individual specific dummy variables is numerically equivalent to OLS on deviations from means. This means that including individual specific dummies yields a consistent estimator of the slope parameters (as n goes to infinity), even though the number of parameters is also going to infinity. Unfortunately, as was pointed in the classic paper by Neyman and Scott (1948), it is generally not the case that the maximum likelihood estimator will retain its nice asymptotic properties when the number of parameters is allowed to increase with sample size. This is for example seen by considering the maximum likelihood estimator of the variance in a static linear panel data model with normal errors: because the maximum likelihood estimator does not make the degrees–of–freedom correction, it will be inconsistent if the number of parameters is of order n.

Conditional maximum likelihood estimation is a method which, when it is applicable, can be used to construct consistent estimators of panel data models in the presence of individual specifc effects. The idea is as follows. Suppose that a random variable, y_{it}, has distribution $f(\cdot;\theta,\alpha_i)$ where θ is the parameter of interest and is common for all i, whereas α_i is a nuicance parameter which is allowed to differ across i. A sufficient statistic, T_i, for α_i is a function of the data such that the distribution of the data given T_i does not depend on α_i. However, it might well depend on θ. If that is the case, then one can estimate θ by maximum likelihood using the conditional distribution of the data given the sufficient statistics. Andersen (1970) proved that the resulting estimator is consistent and asymptotically normal under appropriate regularity conditions. In the two subsections below, we give examples of how the conditional maximum likelihood estimator can be used to construct estimators of the panel data logit and the panel data poisson regression models.

The problem with conditional maximum likelihood estimation as a general prescription for constructing estimators of nonlinear panel data models is that it is not always possible to find sufficient statistics such that the conditional distribution of the data conditional on the sufficient statistic will depend on θ. This is the case for many of the nonlinear models used in econometrics.

5.1 Conditional Maximum Likelihood Estimation of Logit Models

The simplest interesting nonlinear model for which the conditional likelihood approach works, is the "textbook" logit model studied in Rasch (1960, 1961). With two time periods and an individual specific constant we have,

$$y_{it} = 1\left\{x_{it}\beta + \alpha_i + \varepsilon_{it} \geq 0\right\} \qquad t = 1, 2 \qquad i = 1, ..., n$$

where ε_{i1} and ε_{i2} are independent and logistically distributed, conditional on α_i, x_{i1}, x_{i2}. It follows that

$$\Pr\left(y_{it} = 1 | x_{i1}, x_{i2}, \alpha_i\right) = \frac{\exp\left(x_{it}\beta + \alpha_i\right)}{1 + \exp\left(x_{it}\beta + \alpha_i\right)} \tag{94}$$

In this case it is easy to see how the conditional likelihood approach "eliminates" the individual specific effect. Define events A and B by $A = \{y_{i1} = 0, y_{i2} = 1\}$ and $B = \{y_{i1} = 1, y_{i2} = 0\}$. It is then an easy exercise to show that

$$\Pr\left(y_{i1} = 0, y_{i2} = 1 |\, y_{i1} + y_{i2} = 1, x_{i1}, x_{i2}, \alpha_i\right) \;= \tag{95}$$

$$\Pr\left(A|A\cup B,x_{i1},x_{i2},\alpha_i\right) = \frac{1}{1+\exp\left(\left(x_{i1}-x_{i2}\right)\beta\right)}$$

In words, if we restrict the sample to the observations for which y_{it} changes, then the individual specific effects do not enter the distribution of (y_{i1},y_{i2}) given (x_{i1},x_{i2},α_i) and the distribution of y_{i1} given (x_{i1},x_{i2}) has the form of a logit model with explanatory variable $x_{i1}-x_{i2}$ and coefficient β. Intuitively, the implication is that if we restrict the sample to the observations for which y_{it} changes over time, then β can be estimated by estimating a logit in the restricted sample without having to specify the distribution of the individual specific effects. In a sense, conditioning on $y_{i1}+y_{i2}=1$ has the same effect as differencing the data in a linear panel data model.

More generally, if there are $T>2$ observations for each individual, the conditional distribution of $(y_{i1},...,y_{it})$ given $\sum_{t=1}^T y_{it}$ is

$$P\left(y_{i1},...,y_{it}|\sum_{t=1}^{T}y_{it},x_{i1},...,x_{it},\alpha_i\right) = \frac{\exp\left(\sum_{t=1}^T y_{it}x_{it}\beta\right)}{\sum_{(d_1,...,d_t)\in B}\exp\left(\sum_{t=1}^T d_t x_{it}\beta\right)} \tag{96}$$

where B is the set of all sequences of zeros and ones that have $\sum_{t=1}^T d_{it} = \sum_{t=1}^T y_{it}$. Formally this means that $\sum_{t=1}^T y_{it}$ is a sufficient statistic for α_i, and the implication is that one can used (96) to estimate β. Chamberlain (1980) generalized (96) by deriving the conditional likelihood for the multinomial logit model.

When T is large, the number of terms in the denominator of (96) will be large, and and it can be computationally burdensome to calculate the conditional maximum likelihood estimator. In that case one can estimate β by applying the logic leading to (95) to all pairs of observations for a given individual. In other words, one can maximize

$$\sum_{i=1}^{n}\left(\sum_{s<t}\log\left(\frac{\exp\left(y_{it}\left(x_{it}-x_{is}\right)\beta\right)}{1+\exp\left(\left(x_{it}-x_{is}\right)\beta\right)}\right)\right)$$

Unless $T=2$, this objective function is not a (log–)likelihood, and it will generally be less efficient than the conditional maximum likelihood estimator. The asymptotic distribution of the estimator can be found by noting that it is an extremum estimator.

5.2 Poisson Regression Models

The Poisson regression model with individual specific constants provides another example in which the conditional maximum likelihood estimator can be used. This is a special case of the multiplicative model discussed earlier. For simplicity, consider the case where there are two observations for each individual:

$$y_{it} \sim \text{po}\left(\exp(\alpha_i + x_{it}\beta)\right) \qquad t = 1, 2 \qquad i = 1, \ldots, n. \tag{97}$$

One way to understand why the conditional likelihood approach will work in this model, is to recall that if two independent random variables are both Poisson distributed with means μ_1 and μ_2, respectively, then the distribution of one of them given the sum has a binomial distribution with probability parameter $\frac{\mu_1}{\mu_1+\mu_2}$ and trial parameter given by the sum of the two random variables. It therefore follows that if y_{i1} and y_{i2} are drawn from (97) and we restrict attention to the observations for which $y_{i1} + y_{i2} = K$ (say), then $y_{i1} \sim \text{bi}\left(K, \frac{\exp(x_{i1}\beta)}{\exp(x_{i1}\beta)+\exp(x_{i2}\beta)}\right)$. Since this distribution does not involve the individual specific effects, it can be used to make inference about β. For example, one could estimate β by maximizing

$$L = \sum_i -y_{i1} \ln\left(1 + \exp((x_{i2} - x_{i1})b)\right) - y_{i2} \ln\left(1 + \exp((x_{i1} - x_{i2})b)\right)$$

(see, for example, Hausman, Hall and Griliches (1984)).

Recent papers by Blundell, Griffith and Windmeijer (1997), and Lancaster (1997) have pointed out that for the Poisson regression model, (97), the conditional maximum likelihood estimator is identical to the maximum likelihood estimator of β based on maximizing the likelihood function for (97) over b and all the individual specific effects, α_i.

6 Discrete Choice Models with “Fixed” Effects

Manski (1987) made the first successful attempt of consistently estimating a nonlinear panel data model with individual specific “fixed” effects in a situation in which the conditional maximum likelihood approach cannot be applied. His estimator is based on the maximum score estimator (see Manski (1975)) for the binary choice model

$$y_i = 1\left\{x_i\beta + \varepsilon_i \geq 0\right\} \tag{98}$$

Since $P\left(y_i = 1 \middle| x_i\right) = F_{-\varepsilon_i|x_i}\left(x_i\beta\right)$ it follows that if Median$(\varepsilon_i|x_i) = 0$ (uniquely), then observations with $x_i\beta > 0$ will have probabilities greater than $\frac{1}{2}$ and observations with $x_i\beta < 0$ will have probabilities less than $\frac{1}{2}$. In other words,

$$\operatorname{sgn}\left(\Pr\left(y_i = 1|x_i\right) - \Pr\left(y_i = 0|x_i\right)\right) = \operatorname{sgn}(x_i\beta)$$

Under mild regularity conditions, this implies that $E\left[\operatorname{sgn}\left(2y_i - 1\right)\operatorname{sgn}\left(x_ib\right)\right]$ is uniquely maximized at $b = \beta$, and the analogy principle therefore suggests estimating β by

$$\hat{\beta} = \arg\max_b \sum_{i=1}^{n} \operatorname{sgn}\left(2y_i - 1\right)\operatorname{sgn}\left(x_ib\right)$$

Under mild conditions, this estimator is consistent (see Manski, 1985), but it does not converge at rate root–n and it is not asymptotically normal (see Cavanagh(1987) and Kim and Pollard (1990)).[20]

The insight behind Manski's (1987) estimator of the ("non–logit") binary choice model with individual specific effects, is that under mild conditions, exactly the same conditioning that leads from the logit model with individual specific fixed effects (94) to a logit model without the individual specific "fixed" effects, (95), will also lead from the model

$$y_{it} = 1\left\{x_{it}\beta + \alpha_i + \varepsilon_{it} \geq 0\right\} \qquad t = 1, 2; \quad i = 1, ..., n \tag{99}$$

to a model in which the maximum score estimator can be applied. The key assumption is that the distribution of ε_{it} is stationary, in the sense that ε_{i1} and ε_{i2} are identically

[20] Under assumptions that are slightly stronger than Manski's, Horowitz (1992) proposed a smoothed version of the maximum score estimator which does have an asymptotic normal distribution, although the rate is, again, slower than root—n. The rate of convergence of Horowitz's estimator depends on the assumed degree of smoothness of the distribution of the explanatory variables.

distributed conditional on $(x_{i1}, x_{i2}, \alpha_i)$. With this assumption, Manski showed that

$$\Pr\left(y_{i2} = 1 | x_{i1}, x_{i2}, y_{i1} + y_{i2} = 1\right) \lesseqqgtr 1/2$$

depending on whether

$$\left(x_{i2} - x_{i1}\right)\beta \lesseqqgtr 0.$$

The intuition for this result is simple. If the distribution of $-\varepsilon_{i1}$ (and $-\varepsilon_{i2}$) for individual i is $F_i(\cdot)$, then the probability that $y_{it} = 1$ for individual i is $F_i\left(x_{it}\beta + \alpha_i\right)$; this means that for a given individual, higher values of $x_{it}\beta$ are more likely to be associated with $y_{it} = 1$.

Mimicking Manski (1975), this suggests a conditional maximum score estimator defined by

$$\hat{\beta} = \arg\max_b \sum_{i=1}^{n} \operatorname{sgn}\left(y_{i2} - y_{i1}\right) \operatorname{sgn}\left((x_{i2} - x_{i1})b\right) \tag{100}$$

If the panel is of length longer that T, one can estimate β by considering all pairs of observations

$$\hat{\beta} = \arg\max_b \sum_{i=1}^{n} \sum_{s<t} \left(\operatorname{sgn}\left(y_{is} - y_{it}\right) \operatorname{sgn}\left((x_{is} - x_{it})b\right)\right) \tag{101}$$

As was the case for the cross sectional maximum score estimator, this estimator will be consistent under mild regularity conditions. In particular, compared to the logit model considered earlier, it not only leaves the distribution of the errors unspecified, but it also allows for general serial correlation and heteroskedasticity across individuals (but not over time). However, the estimator is not root–n consistent, and not asymptotically normal.[21]

Since on one hand, Manski's estimator is not root–n consistent, but makes very weak assumptions on the errors, and on the other hand assuming a logistic distribution on the errors leads to a root–n consistent and asymptotically normal estimator, it is natural to

[21] Kyriazidou (1995) and Charlier, Melenberg and van Soest (1995) have shown that the same trick used by Horowitz (1992) to modify the maximum score estimator can be used to modify the conditional maximum score estimator. This results in a smoothed conditional maximum score estimator which does have an asymptotic normal distribution, although the rate is, again, slower than root$-n$.

ask whether there are alternative assumptions on the errors that lead to a situation where it is possible to estimate the β– vector at the usual root–n rate. Perhaps surprisingly, the answer to that question seems to be negative. Subject to weak regularity conditions Chamberlain (1993), showed that even if ε_{it} in (99) are i.i.d. with known distribution and independent of $(x_{i1}, x_{i2}, \alpha_i)$, β can be estimated root–n consistently only in the logit case.

It is clear that scale normalizations are needed in each period in order for β in (99) to be identified. Both the logit version of (99) and Manski's treatment impose such scale normalizations. In the logit case, this normalization comes from the variance of the logistic distribution. In Manski's case it is through a scale normalization on β and through the assumption that the errors are identically distributed in the two time periods. In addition to these scale normalizations, the estimators of (99) also assume that the effect of the fixed effect is the same in the two periods. This is in contrast to the linear model in which it is possible to estimate time specific coefficients (factor loadings) on the fixed effect. It is clear that the logic behind the two estimators of the binary choice panel data model discussed here would break down with such factor loadings, but it is less clear whether they would make the model unidentified.

7 Tobit-Type Models with "Fixed" Effects

7.1 Censored Regression Models

The censored regression model is given by

$$\begin{aligned} y_i^* &= x_i\beta + \varepsilon_i \qquad (102) \\ y_i &= \max\{y_i^*, c\} \end{aligned}$$

In text–book treatments, c is usually 0. Note that for $c = -\infty$, (102) becomes the linear regression model, and that one can change the max to a min by a simple change of sign. The censored regression model has been used in many different contexts. In some, c is the lowest possible value that some economic variable can take, and y^* is the desired level of that variable in the absence of this constraint. In other cases, the censoring is induced by the way the data is constructed. For example, earnings variables are sometimes top–coded for confidentiality reasons.

In a panel data context, the censored regression model may be described by

$$\begin{aligned} y_{it}^{*} &= x_{it}\beta + \alpha_i + \varepsilon_{it} \\ y_{it} &= \max\left\{y_{it}^{*}, c\right\} \end{aligned} \tag{103}$$

This model was introduced by Heckman and MaCurdy (1980) in the context of female labor supply.

Because the individual specific effect α_i does not enter linearly or multiplicatively, it is not possible to "difference" it out as was the case for the linear regression model, and it is also unclear under what conditions a conditional likelihood approach can be used to eliminate α_i. Honoré (1992) proposed a different approach to estimating β in this model. The motivation for the estimators given below is different from that in Honoré (1992) because we want to motivate a larger class of estimators. Honoré (1992) also considered estimation of the truncated version of the model. The latter is less interesting and will not be discussed here.

The idea behind the estimator in Honoré (1992) is to artificially censor the dependent variable in such a way that the individual specific effect can be differenced away. This is similar to the approach in Powell (1986) who artificially censored the dependent variable in a cross sectional censored regression model, in such a way that the moment conditions for OLS apply. Specifically, one can define pairs of "residuals" that depend on the individual specific effect in exactly the same way. Intuitively, this implies that differencing the residuals will eliminate the fixed effects.

Define

$$v_{ist}\left(b\right) = \max\left\{y_{is}, c + \left(x_{is} - x_{it}\right) b\right\} - \max\left\{c, c + \left(x_{is} - x_{it}\right) b\right\}$$

At $b = \beta$, we have

$$\begin{aligned} v_{ist}\left(\beta\right) &= \max\left\{y_{is}, c + \left(x_{is} - x_{it}\right)\beta\right\} - \max\left\{c, c + \left(x_{is} - x_{it}\right)\beta\right\} \\ &= \max\left\{\alpha_i + \varepsilon_{is}, c - x_{is}\beta, c - x_{it}\beta\right\} - \max\left\{c - x_{is}\beta, c - x_{it}\beta\right\} \end{aligned}$$

The key observation is that $v_{ist}(\beta)$ is symmetric in s and t. Therefore, if ε_{it}, $t = 1, ..., T$, are independent and identically distributed conditional on (x_i, α_i), where x_i denotes all the explanatory variables for individual i, then $v_{ist}(\beta)$ and $v_{its}(\beta)$ are independent and identically distributed (conditional on (x_i, α_i)). This means that any function of $v_{ist}(\beta)$ minus the same function of $v_{its}(\beta)$ will be symmetrically distributed around 0. We therefore have the conditional moment condition

$$E\left[\left(\xi\left(\psi\left(v_{its}(\beta)\right) - \psi\left(v_{ist}(\beta)\right)\right)\right)\mid x_i, \alpha_i\right] = 0 \tag{104}$$

for any increasing function $\psi(\cdot)$ and any increasing and odd function $\xi(\cdot)$, provided that the expectations are well–defined. The reason why $\psi(\cdot)$ and $\xi(\cdot)$ are assumed to be increasing will become clear shortly.

One could in principle consider estimation of β on the basis of (104). One problem with this is that although β satisfies (104), it does not follow from the previous discussion that there are no other values of the parameter that also satisfy (104). However (104) implies

$$E\left[\left(\xi\left(\psi\left(v_{its}(\beta)\right) - \psi\left(v_{ist}(\beta)\right)\right)\right)(x_{it} - x_{is})\right] = 0 \tag{105}$$

which has the form

$$E\left[r\left(y_{is}, y_{it}, (x_{is} - x_{it})\beta\right)(x_{is} - x_{it})\right] = 0 \tag{106}$$

where $r(\cdot, \cdot, \cdot)$ is a monotone function of its third argument, because of the assumption that $\psi(\cdot)$ and $\xi(\cdot)$ are increasing[22]. By integrating $r(\cdot)$ with respect to its third argument, one can typically turn (106) into the first order condition for a convex minimization problem of the form

$$\min_b E\left[R\left(y_{is}, y_{it}, (x_{is} - x_{it})b\right)\right]. \tag{107}$$

[22] $v_{ist}(b) = \max\{y_{is}, c + (x_{is} - x_{it})b\} - \max\{c, c + (x_{is} - x_{it})b\}$ is monotone in $(x_{is} - x_{it})b$ because $y_{is} \geq c$. It therefore follows that $\xi(\psi(v_{its}(b)) - \psi(v_{ist}(b)))$ depends on b only through $(x_{is} - x_{it})b$ and that it is monotone in $(x_{is} - x_{it})b$.

The parameter β can then be estimated by minimizing a sample analog of (107). It follows from standard results about extremum estimators that the resulting estimator will be consistent and root-n asymptotically normal.

For example, with $\xi(d) = \psi(d) = d$, $c = 0$ and $T = 2$, the function to be minimized in (107) becomes

$$\begin{aligned} &E\left[(\max\{y_{i1}, \triangle x_i b\} - \max\{y_{i2}, -\triangle x_i b\} - \triangle x_i b)^2\right. \\ &+2 \cdot 1\{y_{i1} < \triangle x_i b\}(\triangle x_i b - y_{i1})y_{i2} + 2 \cdot 1\{y_{i2} < -\triangle x_i b\}(-\triangle x_i b - y_{i2})y_{i1}] \end{aligned}$$

which suggests estimating β by

$$\begin{aligned} \widehat{\beta} &= \arg\min_b \sum_{i=1}^{n} (\max\{y_{i1}, \triangle x_i b\} - \max\{y_{i2}, -\triangle x_i b\} - \triangle x_i b)^2 \\ +2 \cdot 1\{y_{i1} &< \triangle x_i b\}(\triangle x_i b - y_{i1})y_{i2} + 2 \cdot 1\{y_{i2} < -\triangle x_i b\}(-\triangle x_i b - y_{i2})y_{i1} \end{aligned}$$

Letting $\xi(d) = sign(d)$ and $\psi(d) = d$, results in the estimator

$$\begin{aligned} \widehat{\beta} &= \arg\min_b \sum_{i=1}^{n} (1 - 1\{y_{i1} \leq \triangle x_i b, y_{i2} \leq 0\}) \\ &\quad \cdot (1 - 1\{y_{i2} \leq -\triangle x_i b, y_{i1} \leq 0\}) \left|y_{i1} - y_{i2} - \triangle x_i b\right| \end{aligned}$$

These are the estimators discussed in detail in Honoré (1992). Honoré and Kyriazidou (1999) discuss estimators defined by a general $\psi(d)$ and $\xi(d) = d$ as well as $\psi(d) = d$ and general $\xi(d)$. The case with panels of length $T > 2$ can be dealt with by considering all pairs of time periods, s and t, as in (101)

The moment condition, (105), was derived from the assumption that ε_{is} and ε_{it} are independent and identically distributed conditional on (x_i, α_i). This assumption is stronger than necessary. To see why, assume the conditional exchangeability assumption that $(\varepsilon_{is}, \varepsilon_{it})$ is distributed like $(\varepsilon_{it}, \varepsilon_{is})$ conditional on (x_i, α_i). This implies that $(\psi(v_{ist}(\beta)), \psi(v_{its}(\beta)))$ is distributed like $(\psi(v_{its}(\beta)), \psi(v_{ist}(\beta)))$, which in turn implies that $\psi(v_{ist}(\beta)) - \psi(v_{its}(\beta))$ is symmetrically distributed around 0 (all conditional on (x_i, α_i)). The moment condition, (105) then follows.

The exchangeability condition is useful because it yields symmetry of $\psi(v_{ist}(\beta)) - \psi(v_{its}(\beta))$, which then yields the moment condition for *any* choice of the odd function,

ξ. On the other hand, if ξ is the identity function, then the moment condition follows if $\psi(v_{ist}(\beta))$ is distributed like $\psi(v_{its}(\beta))$, which is implied by ε_{is} and ε_{it} being identically distributed. In other words, the stationarity assumption that was the key to Manski's estimator for the panel data binary choice model, is also the key to the class of estimators for the panel data censored regression model based on the moment condition (106) (and the minimization problem (107)) with $\xi(d) = d$, whereas the larger class of estimators based on (106) with general ξ seems to require the stronger assumption that ε_{is} and ε_{it} are exchangeable.

7.2 Type 2 Tobit Model (Sample Selection Model)

Kyriazidou (1997) studied the more complicated model

$$\begin{aligned} y_{1it}^{*} &= x_{1it}\beta_1 + \alpha_{1i} + \varepsilon_{1it} \\ y_{2it}^{*} &= x_{2it}\beta_2 + \alpha_{2i} + \varepsilon_{2it} \end{aligned}$$

where we observe:

$$y_{1it} = 1\{y_{1it}^{*} > 0\} \tag{108}$$

$$y_{2it} = \begin{cases} y_{2it}^{*} & \text{if } y_{1it} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{109}$$

This is a panel data version of the sample selection model that Amemiya (1985) calls the Type 2 Tobit Model.

It is clear that β_1 can be estimated by one of the methods for estimation of discrete choice models with individual specific effects discussed earlier. Kyriazidou's insight into estimation of β_2 combines insights from the literature on estimation of semiparametric sample selection models with the idea of eliminating the individual specific effects by first–differencing the data. Specifically, to difference out the individual specific effects α_{2i}, one must restrict attention to observations for which y_{2it}^{*} is observed. With this "sample selection", the mean of the error term in period t is

$$\lambda_{it} = E\left(\varepsilon_{2it} \middle| \varepsilon_{1it} > -x_{1it}\beta_1 - \alpha_{1i}, \varepsilon_{1is} > -x_{1is}\beta_1 - \alpha_{1i}, \zeta_i\right)$$

where $\zeta_i = (x_{1is}, x_{2is}, x_{1it}, x_{2it}, \alpha_{i1}, \alpha_{i2})$. The key observation in Kyriazidou (1997) is that if $(\varepsilon_{1it}, \varepsilon_{2it}, \varepsilon_{1is}, \varepsilon_{2is})$ and $(\varepsilon_{1is}, \varepsilon_{2is}, \varepsilon_{1it}, \varepsilon_{2it})$ are identically distributed (conditional on $(x_{1is}, x_{2is}, x_{1it}, x_{2it}, \alpha_{i1}, \alpha_{i2})$), then for an individual, i, who has $x_{1it}\beta_1 = x_{1is}\beta_1$,

$$\begin{aligned}
\lambda_{it} &= E\left(\varepsilon_{2it} | \, \varepsilon_{1it} > -x_{1it}\beta_1 - \alpha_{1i}, \varepsilon_{1is} > -x_{1is}\beta_1 - \alpha_{1i}, \zeta_i\right) \qquad (110)\\
&= E\left(\varepsilon_{2is} | \, \varepsilon_{1is} > -x_{1is}\beta_1 - \alpha_{1i}, \varepsilon_{1it} > -x_{1it}\beta_1 - \alpha_{1i}, \zeta_i\right)\\
&= \lambda_{is}.
\end{aligned}$$

This implies that for individuals with $x_{1it}\beta_1 = x_{1is}\beta_1$, the same first differencing that will eliminate the fixed effect will also eliminate the efect of sample selection. This suggests a two–step estimation procedure similar to Heckman's (1976, 1979) two–step estimator of sample selection models: first estimate β_1 by one of the methods discussed earlier, and then, secondly, estimate β_2 by applying OLS to the first differences, but giving more weight to observations for which $(x_{1it} - x_{1is})\,\hat{\beta}_1$ is close to zero:

$$\begin{aligned}
\hat{\beta}_2 &= \left[\sum_{i=1}^{n}\sum_{s<t}(x_{2it} - x_{2is})'(x_{2it} - x_{2is})\,K\left(\frac{(x_{1it} - x_{1is})\,\hat{\beta}_1}{h_n}\right) y_{1it}y_{1is}\right]^{-1}\\
&\quad \times \left[\sum_{i=1}^{n}\sum_{s<t}(x_{2it} - x_{2is})'(y_{2it} - y_{2is})\,K\left(\frac{(x_{1it} - x_{1is})\,\hat{\beta}_1}{h_n}\right) y_{1it}y_{1is}\right]
\end{aligned}$$

where K is a kernel and h_n is a bandwidth which shrinks to zero as the sample size increases. Kyriazidou showed that the resulting estimator is $\sqrt{nh_n}$–consistent and asymptotically normal.

Kyriazidou estimator is closely related to the estimator proposed by Powell (1987). That paper considered a cross sectional sample selection model and applied the argument leading to (110) to all pairs of observations, i and j.

7.3 Other Tobit–type Models

As pointed out in Honoré and Kyriazidou (1999), the estimators proposed in Honoré (1992) and Kyriazidou (1997) can be modified fairly trivially to cover the other Tobit–type models discussed in Amemiya (1985). Consider for example, the Type 3 Tobit model with individual–specific effects,

$$y_{1it}^{*} = x_{1it}\beta_1 + \alpha_{1i} + \varepsilon_{1it}$$

$$y_{2it}^{*} = x_{2it}\beta_2 + \alpha_{2i} + \varepsilon_{2it}$$

$$y_{1it} = \begin{cases} y_{1it}^{*} & \text{if } \; y_{1it}^{*} > 0 \\ 0 & \text{if } \; y_{1it}^{*} \leq 0 \end{cases}$$

$$y_{2it} = \begin{cases} y_{2it}^{*} & \text{if } \; y_{1it}^{*} > 0 \\ 0 & \text{if } \; y_{1it}^{*} \leq 0 \end{cases}$$

In that model, the event

$$E = \{y_{1is} > \max\{0, (x_{1is} - x_{1it})\beta_1\},\, y_{1it} > \max\{0, (x_{1it} - x_{1is})\beta_1\}\}$$

is the same as the event

$$\begin{aligned} &\big\{\varepsilon_{1is} > \max\{-x_{1is}\beta_1 - \alpha_{1_i}, -x_{1it}\beta_1 - \alpha_{1_i}\}, \\ &\varepsilon_{1it} > \max\{-x_{1is}\beta_1 - \alpha_{1_i}, -x_{1it}\beta_1 - \alpha_{1_i}\}\big\} \end{aligned}$$

With the exchangeability assumption that $(\varepsilon_{1it}, \varepsilon_{2it}, \varepsilon_{1is}, \varepsilon_{2is})$ and $(\varepsilon_{1is}, \varepsilon_{2is}, \varepsilon_{1it}, \varepsilon_{2it})$ are identically distributed (conditional on $(x_{1is}, x_{2is}, x_{1it}, x_{2it}, \alpha_{i1}, \alpha_{i2})$)

$$\varepsilon_{1is} - \varepsilon_{1it} = (y_{2is} - y_{2it}) - (x_{2is} - x_{2it})\beta_2$$

is symmetrically distributed around 0 conditional on E and conditional on $(x_{1is}, x_{2is}, x_{1it}, x_{2it}, \alpha_{i1}, \alpha_{i2})$ This suggests a two-step approach, where the first step is estimation of β_1 by one of the estimators of the panel data censored regression, and the second step is estimates β_2 by

$$\begin{aligned} \widehat{\beta}_2 \;=\; \arg\min_b \sum_i \sum_{s<t} 1 \Big\{ & y_{1is} > \max\{0, (x_{1is} - x_{1it})\hat{\beta}_1\}, \\ & y_{1it} > \max\{0, (x_{1it} - x_{1is})\hat{\beta}_1\}\Big\} \cdot \Xi\left((y_{is} - y_{it}) - (x_{is} - x_{it})\, b\right) \end{aligned}$$

where Ξ is some symmetric loss function such as $\Xi(d) = d^2$ or $\Xi(d) = |d|$.

The Type 3 Tobit model was also considered by Ai and Chen (1992) who presented moment conditions similar to those implied by the two–step estimator above, although they derived their conditions under the assumption that the errors are independent over time.

It is also straightforward to consider panel data versions of Amemiya's Type 4 and Type 5 Tobit Models. Let

$$\begin{aligned} y_{1it}^{*} &= x_{1it}\beta_1 + \alpha_{1i} + \varepsilon_{1it} \\ y_{2it}^{*} &= x_{2it}\beta_2 + \alpha_{2i} + \varepsilon_{2it} \\ y_{3it}^{*} &= x_{3it}\beta_3 + \alpha_{3i} + \varepsilon_{3it} \end{aligned}$$

In the Type 4 Tobit model we observe $(y_{1it}, y_{2it}, y_{3it})$ from:

$$y_{1it} = \max\{0, y_{1it}^{*}\} \tag{111}$$

$$y_{2it} = \begin{cases} y_{2it}^{*} & \text{if } y_{1it}^{*} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{112}$$

$$y_{3it} = \begin{cases} y_{3it}^{*} & \text{if } y_{1it}^{*} \leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{113}$$

and we can estimate the parameters of this model by considering (111) and (112) as on Type 3 Tobit model and (111) and (113) as a Type 2 sample selection model.

In the Type 5 Tobit model we observe $(y_{1it}, y_{2it}, y_{3it})$ from:

$$y_{1it} = 1\{y_{1it}^{*} > 0\} \tag{114}$$

$$y_{2it} = \begin{cases} y_{2it}^{*} & \text{if } y_{1it} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{115}$$

$$y_{3it} = \begin{cases} y_{3it}^{*} & \text{if } y_{1it} = 0 \\ 0 & \text{otherwise} \end{cases} \tag{116}$$

and we can treat the two outcome equations (115) and (116) separately and apply Kyriazidou's (1997) estimator to β_2 and β_3.

7.4 Monotone Transformation Models

Estimation of β in the cross sectional linear transformation model,

$$h(y_i) = x_i\beta + \varepsilon_i, \tag{117}$$

has been the topic of a large number of recent papers in econometrics and statistics. In this model, β is often considered the primary parameter of interest with h and the distribution of ε left unspecified except that $h(\cdot)$ is assumed to be monotone and ε independent of x. In some cases, h is assumed to be strictly monotone, whereas other papers do not require this, in which case (117) contains both the binary discrete choice and the censored regression model as special cases. When h is assumed to be strictly monotone, one might think of (117) as a generalization of the Box–Cox model. It is clear that β can only be estimated up to scale, unless a scale normalization is imposed on $h(\cdot)$ or ε. In the following, we will therefore only be concerned with estimation of β up to scale.

In a recent paper, Abrevaya (1999) proposed an estimator of β in a fixed effects version of (117),

$$h_t(y_{it}) = x_{it}'\beta + \alpha_i + \varepsilon_{it} \tag{118}$$

where $h_t(\cdot)$ is assumed strictly increasing. His estimator is similar in spirit to that of Han (1987) for the cross sectional transformation model. The key insight in Abrevaya's paper is to difference across individuals in a given time period, rather than across time periods for a given individual,

$$h_t(y_{it}) - h_t(y_{jt}) = (x_{it} - x_{jt})'\beta + (\alpha_i - \alpha_j) + (\varepsilon_{it} - \varepsilon_{jt}).$$

Because h_t is strictly increasing,

$$\begin{aligned} \Pr(y_{it} &> y_{jt} \mid x_{it}, x_{is}, \alpha_i, x_{jt}, x_{js}, \alpha_j) \qquad (119) \\ &= \Pr(\varepsilon_{jt} - \varepsilon_{it} < (x_{it} - x_{jt})'\beta + (\alpha_i - \alpha_j) \mid x_{it}, x_{is}, \alpha_i, x_{jt}, x_{js}, \alpha_j) \end{aligned}$$

where the motivation for conditioning of the explanatory variables in both time periods t and s, is that we will compare this probability in time period t to the same probability in time period s.

Assume that the errors are stationary (given the explanatory variables in all periods and given the fixed effect). This is the same assumption that was made for the discrete choice model and for the censored regression model. This assumption, combined with random sampling, implies that the distribution of $\varepsilon_{jt} - \varepsilon_{it}$ (given $(x_{it}, x_{is}, \alpha_i, x_{jt}, x_{js}, \alpha_j)$) is the same in the two periods. The right hand side of (119) can then be written as

$F_{ij}\left((x_{it}-x_{jt})'\beta+(\alpha_i-\alpha_j)\right)$. On the other hand, by simple inspection it is clear that

$$\Delta x_i'\beta > \Delta x_j'\beta \Longleftrightarrow (x_{it}-x_{jt})'\beta+(\alpha_i-\alpha_j) > (x_{is}-x_{js})'\beta+(\alpha_i-\alpha_j) \tag{120}$$

where $\Delta x = x_t - x_s$. Combining (119) and (120) we then have[23]

$$\begin{aligned} \Delta x_i'\beta &> \Delta x_j'\beta \Longrightarrow \qquad (121) \\ \Pr(y_{it} &> y_{jt} \mid x_{is}, x_{it}, \alpha_i, x_{js}, x_{jt}, \alpha_j) > \Pr(y_{is} > y_{js} \mid x_{is}, x_{it}, \alpha_i, x_{js}, x_{jt}, \alpha_j) \end{aligned}$$

Equation (121) implies that the function

$$S(b) \equiv E\left[\text{sign}\left((\Delta x_i - \Delta x_j)' b\right)\left(1\left(y_{it} > y_{jt}\right) - 1\left(y_{is} > y_{js}\right)\right)\right] \tag{122}$$

is maximized at $b=\beta$. For the case where there are only two time periods, Abrevaya therefore proposed an estimator defined by maximizing the sample analog of (122),

$$S_n(b) \equiv \binom{n}{2}^{-1} \sum_{i \neq j} \text{sign}((\Delta x_i - \Delta x_j)'b)(1(y_{i2} > y_{j2}) - 1(y_{i1} > y_{j1})) \tag{123}$$

Abrevaya (1999) showed that his estimator is consistent and (root–n) asymptotically normal under appropriate regularity conditions. He also showed that although there are n^2 terms in the sum in (123), it is possible to calculate the sum using $O\left(n \log\left(n\right)\right)$ operations. The computational burden associated with the estimator is therefore much smaller that it appears. The case with $T > 2$ observations for each individual can again be dealt with by considering all pairs of time periods.

Abrevaya (2000) proposed an estimator for a model which is more general than (118). That estimator is based on the same idea as Manski's (1985) maximum score estimator of the panel data binary choice estimator. As is the case for the maximum score estimator, it is possible to show that a smoothed version of Abrevaya's estimator is consistent and asymptotocally normal, although the rate of convergence is slower than root–n.

[23] Some smoothness of the distribution of the errors is needed for the inequality between the probabilities to be strict.

7.5 Nonparametric Regression and Fixed Effects

Porter (1997) introduced individual–specific additive effects in a nonparametric regression model by specifying

$$y_{it} = m_t(x_{it}) + \alpha_i + \varepsilon_{it} \tag{124}$$

where ε_{it} has mean 0 conditional on all (past, current and future) values of the explanatory variables, x_{it}. Porter noted that (124) implies that the conditional mean of $y_{it} - y_{is}$ given (x_{it}, x_{is}) is $\ell(x_{it}, x_{is}) \equiv m_t(x_{it}) - m_s(x_{is})$. The latter can be estimated by standard techniques for nonparametric regression (see e.g. Härdle and Linton (1994)), and $m_t(\cdot)$ can then be recovered (except for an additive constant) by averaging ℓ over its second argument.

7.6 Relationship with Estimators for Some Cross Sectional Models.

The estimators for the panel data versions of the discrete choice model, the censored and truncated regression models, the sample selection model and the monotone transformation model all have "cousins" for the cross sectional versions of the models. The relationship is most easily understood by considering a simple cross sectional linear regression model where the observations consist of *i.i.d.* draws of

$$y_i = \alpha + x_i\beta + \varepsilon_i \tag{125}$$

In this model, any two observations have the same intercept, α. With some potential loss of information, one can therefore think of any two observations as if they are from a (static) linear panel data model with $T = 2$. This suggests forming all pairs of observations, and then estimating the slope–parameters, β, in (125) by

$$\widehat{\beta} = \arg\min_b \sum_{i<j} \left((y_i - y_j) - (x_i - x_j)\, b\right)^2$$

It is an easy exersice to show that this is nothing but the OLS estimator of β in the regression of (125).

The same logic can be applied to nonlinear models. If the model under consideration is such that the parameter β can be estimated from a two–period panel by, say, some minimization problem

$$\widehat{\beta} = \arg\min_b \sum_i g\left(y_{i1}, y_{i2}, x_{i1}, x_{i2}, b\right)$$

then a cross sectional version of the model can be estimated by

$$\widehat{\beta} = \arg\min_b \sum_{i<j} g\left(y_i, y_j, x_i, x_j, b\right).$$

Honoré and Powell (1994) applied this insight to construct estimators for the cross sectional censored and truncated regression models based on the panel data estimators in Honoré (1992).

The panel data estimators for the discrete choice and sample selection models also have cross sectional versions. If Manski's (1987) estimator is applied to all pairs of observations from a cross sectional binary choice model, then the maximum rank correlation estimator of Han (1987) results (although his motivation was quite different and his estimator applies to a more general class of transformations models). Likewise, applying the logic behind Kyriazidou's (1997) estimator of the sample selection model to all pairs of observations in a cross sectional sample selection model results in the estimator proposed by Powell (1987). It is interesting to note that the cross sectional estimator that uses all pairs of observations is root–n consistent in both of these cases, although the corresponding panel data estimator converges at a slower rate.

The situation is a little more complicated for the monotone transformation model because the panel data estimator of that model is itself based on pairwise comparisons across individuals. The cross sectional version that treats each pair of observations as if they came from a panel of lenght 2, is therefore based on comparing pairs of pairs, resulting in an estimator defined by a quadruple sum. This estimator is analyzed in Abrevaya (1999).

Table 3 summarizes the relationship between the panel data estimators and their pairwise comparison counterparts. It also lists the estimator for the cross sectional model which we find to be closest in spirit to the panel data estimator.

Table 3: Relationship between Panel Data Estimators and Pairwise Comparison Estimators

Model	"Motivating" Estimator	Panel Data Estimator	Pairwise Comparison
Discrete Choice	Manski (1975)	Manski (1987)	Han (1987)
Censored Regression	Powell (1986)	Honoré (1992)	Honoré and Powell (1994)
Selection	Powell (1987)	Kyriazidou (1997)	Powell (1987)
Type 3 Tobit		Honoré and Kyriazidou (1999)	Honoré, Kyriazidou and Udry (1997)
Monotone Transformation	Han (1987)	Abrevaya (1999)	Abrevaya (1999)

8 Models with Lagged Dependent Variables

With the exception of the models with multiplicative effects, the non–linear models discussed so far all assume that the explanatory variables are strictly exogenous. This assumption is in sharp contrast to the discussion in the first part of this chapter which focused on linear models with predetemined variables. The assumption of strict exogeneity is important. For example, with two time-periods, the basic idea in the logit model was to consider the probability that $y_{i1} = 1$ conditional on the explanatory variables in both periods and conditional on $y_{i1} \neq y_{i2}$. If the explanatory variables include a lagged dependent variable, then the conditioning set includes y_{i1} and $y_{i1} \neq y_{i2}$. This means that the probability is either 1 or zero and cannot be used to make inference about β. By reviewing each of the other methods described in the previous section, it is clear that the motivation for all of them is based on some statement about the joint distribution of (y_{i1}, y_{i2}) given (x_{i1}, x_{i2}). If the explanatory variable in the second time–period, x_{i2}, includes the lagged dependent variable, y_{i1} then the arguments fail.

In this section, we will review some recently proposed methods for dealing with lagged dependent variables in nonlinear models with fixed effects. It will be seen that some progress has been made in this area, but that the methods that have been proposed are case–specific and often lead to estimators that do not converge at the usual root-n rate. One might conclude from this that it would be more fruitful to take a random effect approach that makes some assumptions on the distribution of the individual–specific effects. However, estimation of dynamic nonlinear models is very difficult even in that case. The main difficulty is the so–called initial conditions problem: if one starts observing the individuals when the process in question is already in progress, then the first observation will depend on the dependent variable in the period before the sample starts. Even if that is observed (or one drops the first observation) one will have to deal with relationship between the first lagged dependent variable and the individual–specific effect. That relationship will depend (in a complicated way) on the parameters of the model, but also on the distribution of the explanatory variables in periods prior to the start of the sample, which is typically unknown. In practice one might "solve" this problem by assuming a flexible functional form for the distribution of the first observation (see for example Heckman (1981b) for a discussion of this approach. One case where one can ignore the initial conditions problem is when one can reasonably assume that the process is observed from the start. For example, if the dependent variable is labor supply and the sample consists of people observed (say) from the time they graduated from high school, then there will be no initial conditions problem.

In the next three sub–sections we discuss some approaches that have been used to generalize the limited dependent variable models discussed earlier to the case where one of the explanatory variables is the lagged dependent variable. Very little is know about how to deal with general predetermined variables in the models that we consider.

8.1 Discrete Choice with State Dependence

Including a lagged dependent variable among the explanatory variables in the discrete choice model with individual specific effects gives the model

$$y_{it} = 1\left\{x_{it}\beta + \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it} \geq 0\right\} \qquad t = 1, ..., T; \qquad i = 1, ..., n \tag{126}$$

In its most general setting, this model allows for three sources of persistance (after controlling for the observed explanatory variable, x) in the event described by y_{it}. Persistance can be the result of serial correlation in the error term, ε, a result of the "unobserved hererogeneity", α, or a result of true state dependence through the term $\gamma y_{i,t-1}$. Distinguishing between these sources of persistance is important in many situations because they have very different policy implications. A policy that temporarily increases the probabality that $y = 1$ will have different implications about future probabilities in a model with true state dependence than in model where the persistance is due to unobserved heterogeneity. See, for example, Heckman (1981a) for a discussion of this. Distinguishing between persistance due to state dependence and due to heterogeneity is also important because they sometimes correspond to different economic models. For example, Chiappori and Salanie (2000) and Chiappori (1998) argue it can be used to distinguish between moral hazard and adverse selection. The pricing system in the French automobile insurance market is such that the incentives for not having an accident are stronger if the driver has had fewer accidents in the past. Thus suggests that accident data should show true state dependence: having an accident this period should lower the probability of an accident next period. On the other hand adverse selection suggests that some drivers are permanantly more likely to have accidents, which corresponds to the individual specific effect α_i in (126).

It is clear that even if the errors are serially independent, the conditions discussed earlier for conditional maximum likelihood estimation of the fixed effects logit model are not satisfied because they implied that ε in time period t is independent of the explanatory variables in time period $t - 1$, — a condition which clearly fails when one of the explanatory variables is the lagged dependent variable. By the same argument, the conditions for the conditional maximum score estimator will not be satisfied in the presence of lagged dependent variable. On the other hand it is also clear that the two sources of persistance in (126) have very different implications. For example consider the case where there are no other explanatory variables: if there is no "state dependence" ($\gamma = 0$) then the sequence $(0, 1, 0, 1)$ would be as likely as the sequence $(0, 0, 1, 1)$. On the other hand if $\gamma < 0$ then the first sequence would be more likely, whereas the second would be more likely if $\gamma > 0$. As pointed out by Heckman (1978), this suggests that one should be able to test for "no state dependence" in a model like (126). As will be seen below, this observation can also be used to estimate γ and β in (126).

Consider first the special case of a logit model where the lagged dependent variable is the *only* explanatory variable,

$$y_{it} = 1\left\{\gamma y_{i,t-1} + \alpha_i + \varepsilon_{it} \geq 0\right\} \qquad t = 1, ..., T; \qquad i = 1, ..., n$$

where ε_{it} is i. i. d., independent of α_i, and logistically distributed. Considering only the first three observations (and the initial condition), we have

$$\Pr\left(y_{it} = 1 | \alpha_i, y_{i0}, ..., y_{i,t-1}\right) = \frac{\exp\left(\gamma y_{i,t-1} + \alpha_i\right)}{1 + \exp\left(\gamma y_{i,t-1} + \alpha_i\right)} \qquad t = 1, 2, 3$$

It is then an easy exercise to see that

$$\Pr\left(y_{i1} = 0 | y_{i1} + y_{i2} = 1, \alpha_i, y_{i0}, y_{i3}\right) = \frac{1}{1 + \exp\left(\gamma\left(y_{i0} - y_{i3}\right)\right)}$$

which does not depend on α_i, and which can therefore be used to make inference on γ (Chamberlian (1978)). More generally, with T observations for each individual, the conditional distribution of of $(y_{i1}, ..., y_{iT})$ given $y_{i1,}$ $\sum_{t=1}^{T} y_{it}$ and y_{iT} is

$$P\left(y_{i1}, ..., y_{iT} \middle| y_{i1}, \sum_{t=1}^{T} y_{it}, y_{iT}, \alpha_i \right) = \frac{\exp\left(\gamma \sum_{t=2}^{T} y_{it} y_{i,t-1}\right)}{\sum_{(d_1,...,d_t)\in B} \exp\left(\gamma \sum_{t=2}^{T} d_t d_{t-1}\right)} \tag{127}$$

where B is the set of all sequences of zeros and ones that have $\sum_{t=1}^{T} d_{it} = \sum_{t=1}^{T} y_{it}$, $d_{i1} = y_{i1}$ and $d_{it} = y_{iT}$. Magnac (1997) presents similar results for the multinomial logit version of this model. He also presents the conditional likelihood function for models with more than one lag.

Honoré and Kyriazidou (2000) modify the calculations leading to the conditional maximum likelihood estimator of a fixed effects logit in such a way that it can be applied to (126). Specifically, assume that ε_{it} in (126) are i.i.d. logistically distributed and that each observation is observed for at least four periods (three periods in which both the exogenous variables and the dependent variable are observed, plus the initial value of y). Unlike the case where the lagged dependent variable is the only explanatory variable, $P\left(y_{i1}, ..., y_{iT} \middle| y_{i0}, \sum_{t=1}^{T} y_{it}, y_{iT}, \{x_{it}\}_{t=1}^{T}, \alpha_i \right)$ will in general depend on α_i, and the conditional likelihood approach will therefore generally break down. However (considering the case with $T = 3$ for simplicity), Honoré and Kyriazidou (2000) showed that

$$P\left(y_{i1}, ..., y_{i3} \middle| y_{i0}, \sum_{t=1}^{3} y_{it}, y_{i3}, \{x_{it}\}_{t=1}^{3}, \alpha_i, x_{i2} = x_{i3} \right) = \frac{1}{1 + \exp((x_{i1} - x_{i2})\beta + \gamma(y_{i0} - y_{i3}))} \tag{128}$$

which does *not* depend on α_i. This suggests estimating β and γ by maximizing a conditional likelihood function based on (128). However of one of the explanatory variables is continuously distributed, there will typically be no observations for which $x_{i1} = x_{i2}$. This is similar to the situation when one wants to estimate a conditional expectation of one random variable given that another takes a particular value. One remedy in that case is to use a kernel estimator to average over observations close to the value. Based on this idea, Honoré and Kyriazidou (2000) estimate γ and β by

$$\begin{aligned}(\hat{\beta}, \hat{\gamma}) &= \underset{(b,g)}{\operatorname{argmax}} \sum_{i=1}^{n} 1\{y_{i1} + y_{i2} = 1\} K\left(\frac{x_{i2} - x_{i3}}{h} \right) \\ &\quad \times \ln\left(\frac{\exp((x_{i1} - x_{i2})b + g(d_{i0} - d_{i3}))^{y_{i1}}}{1 + \exp((x_{i1} - x_{i2})b + g(d_{i0} - d_{i3}))} \right)\end{aligned} \tag{129}$$

where $K(\cdot)$ is a kernel[24] which gives the appropriate weight to observation i, and $h \to 0$ as $n \to \infty$. The main limitation of this approach is that it uses only observations in

[24] The term $K\left(\frac{x_{i2} - x_{i3}}{h}\right)$ in (129) plays the same role as the kernel does in non-parametric regression. In a sample, there will be no two observatiosn for which $x_i = x_j$ if x is continuously distributed. However if the object of interest (typically the conditional expectation) is sufficiently smooth, then we can use observations where x_i is close to x_j, where "close" is defined appropriately. See e.g. Härdle and Linton (1994) for a description of non–parametric regression.

a neighborhood of $x_{i2} = x_{i3}$, so it is necessary to assume that distribution of $x_{i2} - x_{i3}$ to have support in a neighborhood of 0. This rules out time–dummies. Honoré and Kyriazidou (2000) give conditions under which this estimator is consistent and asymptotically normal (although it does not converge at rate root–n), and they discuss generalizations to general T, to multinomial models and to models with more lags.

The same trick as above can be used to modify Manski's conditional maximum score estimator in such a way that it applies to the model

$$y_{it} = 1\left\{x_{it}\beta + \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it} \geq 0\right\} \qquad t = 1,2,3; \qquad i = 1,...,n$$

where ε_{it} is i.i.d. (independent of (α_i, x_i)) with distribution function F. With A and B defined as before, we have

$$\begin{aligned} & \operatorname{sgn}\left(P\left(y_{i2} = 1 \middle|\, y_{i0}, \sum_{t=1}^{3} y_{it}, y_{i3}, \left\{x_{it}\right\}_{t=1}^{3}, \alpha_i, x_{i2} = x_{i3}\right)\right. \\ & \left. - P\left(y_{i1} = 1 \middle|\, y_{i0}, \sum_{t=1}^{3} y_{it}, y_{i3}, \left\{x_{it}\right\}_{t=1}^{3}, \alpha_i, x_{i2} = x_{i3}\right)\right) \\ = \; & \operatorname{sgn}\left((x_{i2} - x_{i1})\beta + \gamma(d_{i3} - d_{i0})\right) \end{aligned}$$

Mimicking the logic in Manski (1987), this means that we can consistenty estimate β and γ up to scale by

$$\begin{aligned} \left(\hat{\beta}, \hat{\gamma}\right) = \; & \arg\max_{(b,g)} \sum_{i=1}^{n} K\left(\frac{x_{i2} - x_{i3}}{h}\right) \operatorname{sgn}\left(y_{i2} - y_{i1}\right) \\ & \cdot \operatorname{sgn}\left((x_{i2} - x_{i1})b + g(d_{i3} - d_{i0})\right) \end{aligned}$$

8.2 Dynamic Tobit Models

We next turn to the possibility of allowing lagged dependent variables to enter the censored regression model considered earlier. Depending on the context, the relevant lagged dependent variable is either the lagged observed variable or the lagged latent (unobserved variable). Here, we consider only the former case. Specifically, assume that

$$y_{it} = \max\left\{0, \alpha_i + x_{it}\beta + \sum_{\ell=1}^{L} \gamma_\ell y_{i,t-\ell} + \varepsilon_{it}\right\}, \qquad t = 1, \ldots, T \quad i = 1, \ldots, n. \qquad (130)$$

Honoré (1993) demonstrated that for this model, it is possible to obtain moment conditions that must be satisfied at the true parameter values. To see how this can be done, assumed that $\gamma_\ell \geq 0$ for $\ell = 1, \ldots, L$, and define "residuals" by

$$v_{ist}(b, g) \equiv \max\left\{0, (x_{it} - x_{is})b, y_{it} - \sum_{\ell=1}^{L} g_\ell y_{i,t-\ell}\right\} - x_{it}b.$$

Then

$$\begin{aligned} v_{ist}(\beta, \gamma) &\equiv \max\left\{0, (x_{it} - x_{is})\beta, y_{it} - \sum_{\ell=1}^{L} \gamma_\ell y_{i,t-\ell}\right\} - x_{it}\beta \\ &= \max\left\{-x_{it}\beta, -x_{is}\beta, \alpha_i + \varepsilon_{it}\right\} \end{aligned}$$

If $\{x_{it}\}_{t=1}^{T}$ is strictly exogenous in the sense that ε_{it} and ε_{is} are identically distributed conditional on $\{x_{it}\}_{t=1}^{T}$ then for any function $\psi(\cdot)$,

$$E\left[\psi\left(v_{ist}(\beta, \gamma)\right) - \psi\left(v_{its}(\beta, \gamma)\right) \middle| \{x_{it}\}_{t=1}^{T}\right] = 0 \tag{131}$$

which suggests that (β, γ) can be estimated by GMM. Honoré and Hu (2000) presents a set of sufficient conditions under which (131) is uniquely satisfied at the true parameter value. The most restrictive assumption is that $x_{it} - x_{is}$ has support in a neighborhood around 0, which rules out time–dummies.

Honoré and Hu (2000) also discuss how a modification of the same idea can be used to construct moment conditions for model with general predetermined explanatory variables, and Hu (2000) shows how to generalize the approach so that it can be used to construct moment conditions for a model in which the lagged variables in (130) are the lagged uncensored variables. This is, for example, the relevant model if the censoring is due to top–coding.

8.3 Dynamic Sample Selection Models

Kyriazidou (1999) generalizes her approach to estimation of

$$\begin{aligned} y_{it}^* &= \rho_0 y_{it-1}^* + x_{it}^* \beta_0 + \alpha_i^* + \varepsilon_{it}^* \\ y_{it} &= d_{it} y_{it}^* \\ d_{it} &= 1\left\{\phi_0 d_{it-1} + w_{it}\gamma_0 + \eta_i - u_{it} \leq 0\right\} \end{aligned}$$

This is the same model that Kyriazidou considered in her (1997) paper, except that the model is now dynamic, with both the dependent variables, y_{it}^* and d_{it} depending on their own lagged value. The key insight is to combine the insights from the dynamic linear panel data models with the insight in Kyriazidou (1997). For simplicity assume that $(\varepsilon_{it}^*, u_{it})$ is i.i.d. over time and independent of all other right hand side variables. Applying the methods discussed in the first part of this chapter to observations for which y_{it}^* is observed in three consequetive periods (so $d_{it} = d_{it-1} = d_{it-2} = 1$), will result in a sample selection bias term which after first differencing has the form $E\left[\varepsilon_{it}^* \middle| u_{it} \geq \phi_0 + w_{it}\gamma_0 + \eta_i\right] - E\left[\varepsilon_{it-1}^* \middle| u_{it-1} \geq \phi_0 + w_{it-1}\gamma_0 + \eta_i\right]$. This sample selection term will be 0 for observations for whom $w_{it}\gamma_0 = w_{it-1}\gamma_0$. The idea therefore is to apply the methods discussed in the first part of this paper augmented by kernel–weights that give more weight to observations for which $w_{it}\widehat{\gamma}$ is close to $w_{it-1}\widehat{\gamma}$, where $\widehat{\gamma}$ is an estimate of γ_0 (using, for example, the method proposed in Honoré and Kyriazidou (2000)).

9 "Random" Effects Models

Since little is known about how to deal with fixed effects in nonlinear models other than the ones discussed above, it is often appealing to make assumptions on the distribution of the individual effects. When the distribution of the error is parameterized completely, then the resulting model is usually refered to as random effects model. As mentioned in the previous section, this appoach is problematic in dynamic model if one does not observe the start of the process. On the other hand, there are no conceptual difficulties in estimating the parameters of a random effects model by maximum likelihood or methods of moments if the explanatory variables are strictly exogenous, and the distribution of the errors, ε_{it}, is specified. The downside is that there might be practical difficulties in implementing these methods, since the likelihood function and the conditional moments will typically involve multivariate integration. In that case, simulation based inference can be extremely useful. See for example Hajivassillou and Ruud (1994) or Keanne

(1994). It is also straightforward to consistently estimate the parameters of certain semiparametric random effects models. Consider for example the censored regression model in Section 7.1. If the errors and the individual specific effects are independent of each other and both are independent of the regressors, then β can be estimated by applying one of the many semiparametric estimators of the censored regression model to the pooled data set consisting of the observations for all i and t. The main complication in that case is that one must correct the variance of the estimator to account for the fact that the observations for a given i are not independent (because they all depend on the same individual–specific effect).

A number of papers propose estimators of models that make assumptions that fall between fixed and random effects models. These papers are motivated by the tradeoff between the difficulties in estimating fixed effects versions of nonlinear models and the fairly strong assumptions that one must make in a random effects approach. As an example, consider the discrete choice model of Section 6. Following Chamberlain (1984), if the individual specific effect, α_i, happens to be of the form $\alpha_i = \sum_{t=1}^{T} x'_{it}\gamma_t + u_i$ where u_i and the transitory errors, ε_{it}, are jointly independent of $(x_{i1}, ..., x_{iT})$ then one can apply an estimator of the semiparametric discrete choice model to the data for each time–period to estimate $(\gamma_1, \gamma_2, ..., \gamma_{t-1}, \gamma_t + \beta, \gamma_{t+1}, ..., \gamma_T)$ up to scale. These can then be combined (via minimum distance) to obtain estimators of $\{\gamma_t\}_{t=1}^{T}$ and β (up to scale). In Chamberlain's example, the ε_{it}'s and the u_i's were assumed to be normally distributed, so the estimation could be done by probit maximum likelihood. Although the functional form assumption made on the individual specific effect makes the model much less general than the fixed effects model, it should be noted that the approach does not require the transitory errors to be homoskedastic over time. This is in contrast to the fixed effects estimators which all assumed some kind of stationarity of the errors.

Newey (1994) considered estimation of the Chamberlain's model but with $\alpha_i = \rho(x_{i1}, ..., x_{iT}) + u_i$ where the function ρ is unknown. If F_t is the cummulative distribution function for $u_i + \varepsilon_{it}$ then

$$P(y_{it} = 1 | x_{i1}, ..., x_{iT}) = F_t(\rho(x_{i1}, ..., x_{iT}) + x_{it}\beta)$$

or

$$F_t^{-1}\left(P\left(y_{it}=1\right|x_{i1},...,x_{iT})\right)=\rho\left(x_{i1},...,x_{iT}\right)+x_{it}\beta \tag{132}$$

When the errors are jointly normally distributed, this implies

$$\begin{aligned}\Phi^{-1}\left(P\left(y_{it}=1\right|x_{i1},...,x_{iT})\right) &= \sqrt{\frac{Var\left[u_i+\varepsilon_{is}\right]}{V\left[u_i+\varepsilon_{it}\right]}}\Phi^{-1}\left(P\left(y_{is}=1\right|x_{i1},...,x_{iT})\right) \\ &\quad +\sqrt{\frac{1}{Var\left[u_i+\varepsilon_{it}\right]}}\left(x_{it}-x_{is}\right)\beta\end{aligned}$$

Since discrete choice models can only be estimated up to scale, one can normalize $Var\left[u_i+\varepsilon_{it}\right]=1$ and then estimate β and $\sqrt{Var\left[u_i+\varepsilon_{is}\right]}$ by regressing a nonparametric estimate of $P\left(y_{it}=1\right|x_{i1},...,x_{iT})$ on a nonparametric estimate of $P\left(y_{is}=1\right|x_{i1},...,x_{iT})$ and on $(x_{it}-x_{is})$. Newey (1994) derived the limiting distribution of this estimator. Chen (1998) generalized the model further by allowing the distribution of the errors u and ε to be unknown. His insight is to note that if one normalizes one of the components (say, the first) of β to be one so $\beta=\begin{pmatrix}1\\ \widetilde{\beta}\end{pmatrix}$ then (132) implies that

$$x_{it}^1=-\rho\left(x_{i1},...,x_{iT}\right)-\widetilde{x}_{it}\widetilde{\beta}+F_t^{-1}\left(P\left(y_{it}=1\right|x_{i1},...,x_{iT})\right)$$

or

$$x_{it}^1-x_{is}^1=-\left(\widetilde{x}_{it}-\widetilde{x}_{is}\right)\widetilde{\beta}+F_t^{-1}\left(P\left(y_{it}=1\right|x_{i1},...,x_{iT})\right)-F_s^{-1}\left(P\left(y_{is}=1\right|x_{i1},...,x_{iT})\right) \tag{133}$$

Here $P\left(y_{it}=1\right|x_{i1},...,x_{iT})$ and $P\left(y_{is}=1\right|x_{i1},...,x_{iT})$ can be estimated nonparametrically and $\widetilde{\beta}$ can be estimated by observing that (133) is a partially linear regression model of the type studied by e.g. Robinson (1988).

The idea of writing the individual specific effect as $\alpha_i=\rho\left(x_{i1},...,x_{iT}\right)+u_i$ where u_i is treated as an error term can also be applied to the other models discussed above. See for example Jacubson (1988) or Charlier, Melenberg and van Soest (2000) for applications of this idea in the context of the censored regression model, and Nijman and Verbeek (1992), Zabel (1992) and Wooldridge (1995) for a discussion of this approach in sample selection models.

In a linear model, there is no loss of generality in making assumptions of the form $\alpha_i = \sum_{t=1}^{T} x'_{it}\gamma_t + u_i$ because one can always interpret $\sum_{t=1}^{T} x'_{it}\gamma_t$ as the projection of α_i on $(x_{i1}, ..., x_{iT})$. Making such an assumption in a non–linear model is much more restrictive. In particular, if $\alpha_i = \rho(x_{i1}, ..., x_{iT}) + u_i$ where u_i is independent of $(x_{i1}, ..., x_{iT})$ for some T then the same assumption will typically not be satisfied for some other T. This means that the model which is estimated (and which is assumed to be true) depends on the number of time–series observations, the econometrician happens to observe.

Other alternatives to the "pure" fixed approach have been proposed. For example, Lee (1999) makes assumptions on the joint distribution of the regressors and the individual specific effects which allow him to construct a maximum rank correlation–type estimator of the static discrete choice panel data model. Honoré and Lewbel (2000) exploit the assumption that one of the regressors is independent of the individual specific effect to construct an estimator of a discrete choice panel data model with predetermined explanatory variables.

10 Concluding Remarks

Our discussion has focused on two of the developments in panel data econometrics since the Handbook chapter by Chamberlain (1984). In the first part of the paper we have reviewed linear panel data models with predetermined variables, and in the second we have discussed methods for dealing with nonlinear panel data models. Unfortunately, the intersection of these two literatures is very small. With the exception of multiplicative models and models where the only source of "predetermined–ness" is lagged dependent variables, almost nothing is known about nonlinear models with general predetermined variables. One step in this direction was taken in Arellano and Carrasco (1996). This is an exciting area for future research.

11 References

Abowd, J.M. and D. Card (1989): “On the Covariance Structure of Earnings and Hours Changes”, *Econometrica*, 57, 411-445.

Abrevaya, J. (1999): “Leapfrog Estimation of a Fixed-Effects Model with Unknown Transformation of the Dependent Variable,” *Journal of Econometrics*, 93, 2, pp. 203-228.

Abrevaya J. (2000): “Rank estimation of a generalized fixed-effects regression model,” *Journal of Econometrics,* 95, 1, pp. 1-23.

Ahn, S. and P. Schmidt (1995): “Efficient Estimation of Models for Dynamic Panel Data”, *Journal of Econometrics*, 68, 5-27.

Ai, C. and C. Chen (1992): “Estimation of a Fixed Effect Bivariate Censored regression Model,” *Economic Letters*, 40, 403–406.

Allenby, G.M. and P.E. Rossi (1999): “Marketing Models of Consumer Heterogeneity”, *Journal of Econometrics*, 89, 57-78.

Alonso-Borrego, C. and M. Arellano (1999): “Symmetrically Normalized Instrumental-Variable Estimation Using Panel Data”, *Journal of Business & Economic Statistics*, 17, 36-49.

Alvarez, J. and M. Arellano (1998): “The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators”, Working Paper 9808, CEMFI, Madrid.

Amemiya, T. (1985): *Advanced Econometrics.* Harvard University Press.

Amemiya, T. and T.E. MaCurdy (1986): “Instrumental-Variable Estimation of an Error-Components Model”, *Econometrica*, 54, 869-881.

Andersen, E. (1970): “Asymptotic Properties of Conditional Maximum Likelihood Estimators,” *Journal of the Royal Statistical Society*, Series B, 32, pp. 283-301.

Anderson, T.W., N. Kunitomo, and T. Sawa (1982): "Evaluation of the Distribution Function of the Limited Information Maximum Likelihood Estimator", *Econometrica*, 50, 4, 1009-1027.

Anderson, T.W. and C. Hsiao (1981): "Estimation of Dynamic Models with Error Components", *Journal of the American Statistical Association*, 76, 598-606.

Anderson, T.W. and C. Hsiao (1982): "Formulation and Estimation of Dynamic Models Using Panel Data", *Journal of Econometrics*, 18, 47-82.

Arellano, M. (1990): "Testing for Autocorrelation in Dynamic Random Effects Models", *Review of Economic Studies*, 57, 127-134.

Arellano, M. (1993): "On the Testing of Correlated Effects with Panel Data", *Journal of Econometrics*, 59, 87-97.

Arellano, M. and S.R. Bond (1988): "Dynamic Panel Data Estimation Using DPD -A Guide for Users", Institute for Fiscal Studies, Working Paper 88/15, London.

Arellano, M. and S.R. Bond (1991): "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations", *Review of Economic Studies*, 58, 277-297.

Arellano, M. and O. Bover (1995): "Another Look at the Instrumental-Variable Estimation of Error-Components Models", *Journal of Econometrics*, 68, 29-51.

Arellano, M. and R. Carrasco (1996): "Binary Choice Panel Data Models with Predetermined Variables," Working Paper 9618, CEMFI, Madrid

Ashenfelter, O. and D. Card (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs", *Review of Economics and Statistics*, 67, 648-660.

Back, K. and D.P. Brown (1993): "Implied Probabilities in GMM Estimators", *Econometrica*, 61, 971-975.

Baltagi, B. H. (1995): *Econometric Analysis of Panel Data.* John Wiley and Sons Ltd.

Baltagi, B., J. Hidalgo, and Q. Li (1996): "A Nonparametric Poolability Test", *Journal of Econometrics*, 75, 345-367.

Bhargava, A. and J.D. Sargan (1983): "Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods", *Econometrica*, 51, 1635-1659.

Becker, G., M. Grossman, and K. Murphy (1994): "An Empirical Analysis of Cigarette Addiction", *American Economic Review*, 84, 396-418.

Bekker, P.A. (1994): "Alternative Approximations to the Distributions of Instrumental Variable Estimators", *Econometrica*, 62, 657-681.

Blundell, R., S. Bond, M.P. Devereux, and F. Schiantarelli (1992): "Investment and Tobin's Q: Evidence from Company Panel Data", *Journal of Econometrics*, 51, 233-257.

Blundell, R., R. Griffith and F. Windmeijer (1997): "Individual Effects and Dynamic Count Data," unpublished manuscript, University College London.

Blundell, R. and S. Bond (1998): "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models", *Journal of Econometrics*, 87, 115-143.

Blundell, R. and S. Bond (1999): "GMM Estimation with Persistent Panel Data: An Application to Production Functions", Working Paper W99/4, Institute for Fiscal Studies, London.

Bond, S. and C. Meghir (1994): "Dynamic Investment Models and the Firm's Financial Policy", *Review of Economic Studies*, 61, 197-222.

Bover, O. (1991): "Relaxing Intertemporal Separability: A Rational Habits Model of Labor Supply Estimated from Panel Data", *Journal of Labor Economics*, 9, 85-100.

Breusch, T.S., G.E. Mizon, and P. Schmidt (1989): "Efficient Estimation Using Panel Data", *Econometrica*, 57, 695-700.

Browning, M. (1992): “Children and Household Economic Behaviour”, *Journal of Economic Literature*, 30, 1434-1475.

Canova, F. and A. Marcet (1995): “The Poor Stay Poor: Nonconvergence Across Countries and Regions”, Economics Working Paper 137, Universitat Pompeu Fabra, Barcelona.

Carrasco, R. (1998): “Binary Choice with Binary Endogenous Regressors in Panel Data: Estimating the Effect of Fertility on Female Labour Participation”, Working Paper 9805, CEMFI, Madrid.

Cavanagh, C. L. (1987): “The Limiting Behavior of Estimators Defined by Optimization” unpublished manuscript, Department of Economics, Harvard University.

Chamberlain, G. (1978): “On the Use of Panel Data”, unpublished manuscript, Department of Economics, Harvard University.

Chamberlain, G. (1980): “Analysis of Covariance with Qualitative Data” *Review of Economic Studies*, 47, 225–238.

Chamberlain, G. (1982a): “Multivariate Regression Models for Panel Data”, *Journal of Econometrics*, 18, 5-46.

Chamberlain, G. (1982b): “The General Equivalence of Granger and Sims Causality”, *Econometrica*, 50, 569-581.

Chamberlain, G. (1984): “Panel Data”, in Griliches, Z. and M.D. Intriligator (eds.) *Handbook of Econometrics*, vol. 2, Elsevier Science, Amsterdam.

Chamberlain, G. (1987): “Asymptotic Efficiency in Estimation with Conditional Moment Restrictions”, *Journal of Econometrics*, 34, 305-334.

Chamberlain, G. (1992a): “Efficiency Bounds for Semiparametric Regression”, *Econometrica*, 60, 567-596.

Chamberlain (1992b): Comment: Sequential Moment Restrictions in Panel Data", *Journal of Business & Economic Statistics*, 10, 20-26.

Chamberlain, G. (1993): "Feedback in Panel Data Models", unpublished manuscript, Department of Economics, Harvard University.

Chamberlain, G. and K. Hirano (1999): "Predictive Distributions Based on Longitudinal Earnings Data", *Annales d'Économie et de Statistique*, 55-56, 211-242.

Charlier, E., B. Melenberg, and A. van Soest (1995): "A Smoothed Maximum Score Estimator for the Binary Choice Panel Data Model and an Application to Labour Force Participation," *Statistica Neerlandica*, 49, pp. 324–342.

Charlier, E., B. Melenberg, and A. van Soest (2000): "Estimation of a Censored Regression Panel Data Model Using Conditional Moments Restrictions Efficiently," *Journal of Econometrics*, 95, 1, pp. 25–56

Chen, S. (1998): "Root–N Consistent Estimation of a Panel Data Sample selection Model" unpublished manuscript, The Hong Kong University of Science and Technology.

Chen, Heckman and Vytlacil (1998): "Identification and $\sqrt{N}$ Estimation of Semiparametric Panel Data Models with Binary Variables and Latent Factors" unpublished manuscript, Department of Economics, University of Chicago.

Chiappori, P.–A. (1998): "Econometric Models of Insurance under Asymmetric Information," unpublished manuscript, Department of Economics, University of Chicago.

Chiappori, P.–A. and B. Salanie (2000): "Testing for Adverse Selection in Insurance Markets," *Journal of Political Economy*, 108, 56–78.

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

Crepon, B., F. Kramarz, and A. Trognon (1997): "Parameters of Interest, Nuisance Parameters and Orthogonality Conditions. An Application to Autoregressive Error Component Models", *Journal of Econometrics*, 82, 135-156.

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

Geweke, J. and M. Keane (2000): "An Empirical Analysis of Earnings Dynamics Among Men in the PSID: 1968-1989", *Journal of Econometrics*, 96, 293-356.

Granger, C.W.J. (1969): "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods", *Econometrica*, 37, 424-438.

Griliches, Z. and J.A. Hausman (1986): "Errors in Variables in Panel Data", *Journal of Econometrics*, 31, 93-118.

Hahn, J. (1997): "Efficient Estimation of Panel Data Models with Sequential Moment Restrictions", *Journal of Econometrics*, 79, 1-21.

Hahn, J. (1998): "Asymptotically Unbiased Inference of Dynamic Panel Data Model with Fixed Effects When Both n and T are Large", Unpublished manuscript, University of Pennsylvania.

Hajivassiliou, V. and D. McFadden (1990): "The Method of Simulated Scores for the Estimation of LDV Models with an Application to External Debt Crisis", Cowles Foundation Discussion Paper 967.

Hajivassilou, V. and P. Ruud (1994): "Classical Estimation Methods for LDV Models using Simulation" in Engle, R.F. and D.L. McFadden (eds.): *Handbook of Econometrics*, vol. 4, Elsevier, Amsterdam, Ch. 40.

Hall, R. and F. Mishkin (1982): "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households", *Econometrica*, 50, 461-481.

Han, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics*, 35, 303-316.

Hansen, L.P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50, 1029-1054.

Hansen, L.P., J. Heaton, and A. Yaron (1996): "Finite Sample Properties of Some Alternative GMM Estimators", *Journal of Business & Economic Statistics*, 14, 262-280.

Härdle, W. and O. Linton (1994): "Applied Nonparametric Methods" in Engle, R.F. and D.L. McFadden (eds.): *Handbook of Econometrics*, vol. 4, Elsevier, Amsterdam, Ch. 38.

Hausman, J.A. and W.E. Taylor (1981): "Panel Data an Unobservable Individual Effects", *Econometrica*, 49, 1377-1398.

Hausman, J.A., B. Hall, and Z. Griliches (1984): "Econometric Models for Count Data with an Application to the Patents-R&D Relationship," *Econometrica*, 52(4), pp. 909–938.

Hayashi, F. and C. Sims (1983): "Nearly Efficient Estimation of Time Series Models with Predetermined, But Not Exogenous, Instruments", *Econometrica*, 51, 783-798.

Hayashi, F. and T. Inoue (1991): "The Relation Between Firm Growth and Q with Multiple Capital Goods: Theory and Evidence from Panel Data on Japanese Firms", *Econometrica*, 59, 731-753.

Heckman, J. J. (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple estimator for Such Models," *Annals of Economic and Social Measurement*, 15, pp. 475–492.

Heckman, J. J. (1978): "Simple Statistical Models for Discrete Panel Data Developed and Applied to Tests of the Hypothesis of True State Dependence against the

Hypothesis of Spurious State Dependence," *Annales de l'INSEE,* 30–31, pp. 227–269.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, pp. 153–161.

Heckman, J. J. (1981a): "Statistical Models for Discrete Panel Data," in *Structural Analysis of Discrete Panel Data with Econometric Applications,* edited by C. F. Manski and D. McFadden.

Heckman, J. J. (1981b): "The Incedental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time–Discrete Data Stochastic Process" in *Structural Analysis of Discrete Panel Data with Econometric Applications,* edited by C. F. Manski and D. McFadden.

Heckman, J.J. (1981c): "Heterogeneity and State Dependence," in *Studies of Labor Markets*, edited by S. Rosen. The National Bureau of Economic Research. The University of Chicago Press.

Heckman, J. J. and T. E. MaCurdy (1980): "A Life Cycle Model of Female Labour Supply." *Review of Economic Studies*, 47, pp. 47–74.

Hillier, G.H. (1990): "On the Normalization of Structural Equations: Properties of Direction Estimators", *Econometrica*, 58, 1181-1194.

Holtz-Eakin, D., W. Newey, and H. Rosen (1988): "Estimating Vector Autoregressions with Panel Data", *Econometrica*, 56, 1371-1395.

Honoré, B. E. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica*, 60, pp. 533-565.

Honoré, B. E. (1993): "Orthogonality Conditions for Tobit Models with Fixed Effects and Lagged Dependent Variables," *Journal of Econometrics*, 59, pp. 35-61.

Honoré, B. E., and L. Hu (1999): "Estimation of Censored Regression Models with Endogeneity" unpublished manuscript, Department of Economics, Princeton University.

Honoré, B. E. and E. Kyriazidou (1999): "Estimation of Tobit–Type Models with Individual Specific Effects," *Econometric Reviews* (forthcoming).

Honoré, B. E. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 4, pp. 839-874.

Honoré, B. E., E. Kyriazidou and C. Udry (1997):"Estimation of Type 3 Tobit Models using Symmetric Trimming and Pairwise Comparisons," *Journal of Econometrics*, 76, pp. 107–128.

Honoré, B. E., and Lewbel (2000): "Semiparametric Binary Choice Panel Data Models Without Strictly Exogeneous Regressors," unpublished manuscript, Department of Economics, Princeton University.

Honoré, B. E., and J. L. Powell (1994): "Pairwise Difference Estimators of Censored and Truncated Regression Models," *Journal of Econometrics*, 64(2)**:** 241–278.

Horowitz, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, pp. 505-531.

Horowitz, J. L. and M. Markatou (1996): "Semiparametric Estimation of Regression Models for Panel Data", *Review of Economic Studies*, 63, 145-168.

Hsiao, C. (1986): *Econometric Analysis of Panel Data.* Cambridge University Press.

Hu, L. (2000): "Estimating a Censored Dynamic Panel Data Model with an Application to Earnings Dynamics," unpublished manuscript, Department of Economics, Northwestern University.

Imbens, G. (1997): "One-step Estimators for Over-identified Generalized Method of Moments Models", *Review of Economic Studies*, 64, 359-383.

Imbens, G., R. Spady, and P. Johnson (1998): “Information Theoretic Approaches to Inference in Moment Condition Models”, *Econometrica*, 66, 333-357.

Jacubson, G (1988): ”The Sensitivity of Labor Supply Parameter Estimates to Unobserved Individual Effects: Fixed and Random Effects Estimates in a Nonlinear Model Using Panel Data,” *Journal of Labor Economics*, 6, 302–329.

Kao, C. (1999): “Spurious Regression and Residual-Based Tests for Cointegration in Panel Data”, *Journal of Econometrics*, 90, 1-44.

Keane, M. (1993): “Simulation Estimation for Panel Data Models with Limited Dependent Variables”, in Maddala, G.S., C.R. Rao and H.D. Vinod (eds.) Handbook of Statistics, Vol. 11, Elsevier Science.

Keane, M. (1994): “A Computationally Practical Simulation Estimator for Panel Data”, *Econometrica*, 62, 95-116.

Keane, M. and D. Runkle (1992): “On the Estimation of Panel-Data Models with Serial Correlation When Instruments Are Not Strictly Exogenous”, *Journal of Business & Economic Statistics*, 10, 1-9.

Kiviet, J.F. (1995): “On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models”, *Journal of Econometrics*, 68, 53-78.

Kunitomo, N. (1980): “Asymptotic Expansions of the Distribution of Estimators in a Linear Functional Relationship and Simultaneous Equations”, *Journal of the American Statistical Society*, 75, 693-700.

Kyriazidou, E. (1995): “Essays in Estimation and Testing of Econometric Models”. Northwestern University, Ph.D. dissertation.

Kyriazidou, E. (1997): “Estimation of a Panel Data Sample Selection Model,” *Econometrica*, 65, pp. 1335–1364.

Kyriazidou, E. (1999): “Estimation of Dynamic Panel Data Sample Selection Models,” unpublished manuscript, Department of Economics, University of Chicago.

Lancaster, T. (1997): “Orthogonal Parameters in Panel Data,” Working Paper No. 97–12, Brown University, Department of Economics.

Lee, M.–J. (1999): “A Root–N Consistent Semiparametric Estimator for Related Effect Binary Response Panel Data,” Econometrica, 67, 427–434.

Li, Q. and T. Stengos (1996): “Semiparametric Estimation of Partially Linear Panel Data Models”, *Journal of Econometrics*, 71, 389-397.

Li, Q. and C. Hsiao (1998): “Testing Serial Correlation in Semiparametric Panel Data Models”, *Journal of Econometrics*, 87, 207-237.

Lillard, L. and R.J. Willis (1978): “Dynamic Aspects of Earnings Mobility”, *Econometrica*, 46, 985-1012.

MaCurdy, T.E. (1982): “The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis”, *Journal of Econometrics*, 18, 83-114.

Magnac, T. (1997): “State Dependence and Heterogeneity in Youth Employment Histories,” Working Paper, INRA and CREST, Paris.

Manski, C. (1975): “The Maximum Score Estimation of the Stochastic Utility Model of Choice”, *Journal of Econometrics*, 3, pp. 205-228.

Manski, C. (1985): “Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator,” *Journal of Econometrics*, 27, pp. 313-333.

Manski, C. (1987): “Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data,” *Econometrica*, 55, pp. 357-362.

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

Morimune, K. (1983): "Approximate Distributions of k-Class Estimators when the Degree of Overidentifiability Is Large Compared with the Sample Size", *Econometrica*, 51, 821-841.

Neyman J., and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations," *Econometrica*, 16, 1–32.

Newey, W. K. (1994): "The Asymptotic Variance of Semiparametric estimators" *Econometrica,* 62, 1349–1382.

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

Nijman, T., and M. Verbeek (1992): "Nonresponse in Panel Data: The Impact on Estimates of a Life Cycle Consumption Function," *Journal of Applied Econometrics*, 7, 243-257.

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

Phillips, P.C.B. (1983): "Exact Small Sample Theory in the Simultaneous Equations Model", in Griliches, Z. and M.D. Intriligator (eds.): *Handbook of Econometrics*, vol. 1, North-Holland, Amsterdam, Ch. 8.

Phillips, P.C.B. and H.R. Moon (1999): "Linear Regression Limit Theory for Nonstationary Panel Data", *Econometrica*, 67, forthcoming.

Kim, J. and D. Pollard (1990): "Cube Root Asymptotics," *Annals of Statistics*, 18, pp. 191-219.

Porter, J. (1997): "Nonparametric Regression Estimation for a Panel data Model with Additive Individual Efects." Harvard University, unpublished.

Powell, J. L. (1984): "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, pp. 303-325.

Powell, J. L. (1986): "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 54, pp. 1435-1460.

Powell, J. L. (1987): "Semiparametric Estimation of Bivariate Latent Models," Working Paper no. 8704, Social Systems Research Institute, University of Wisconsin–Madison.

Powell, J. L. (1994): "Estimation of Semiparametric Models", in Engle, R.F. and D.L. McFadden (eds.): *Handbook of Econometrics*, vol. 4, Elsevier, Amsterdam, Ch. 41.

Qin, J. and J. Lawless (1994): "Empirical Likelihood and General Estimating Equations", *Annals of Statistics*, 22, 300-325.

Rasch, G. (1960): "Probabilistic Models for Some Intelligence and Attainment Tests," Denmarks Pædagogiske Institut, Copenhagen.

Rasch, G. (1961): "On the General Laws and the Meaning of Measurement in Psychology," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol 4, University of California Press, Berkeley and Los Angeles.

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, pp. 931-954.

Runkle, D.E. (1991): "Liquidity Constraints and the Permanent Income Hypothesis: Evidence from Panel Data", *Journal of Monetary Economics*, 97, 73-98.

Sargan, J.D. (1958): "The Estimation of Economic Relationships Using Instrumental Variables", *Econometrica*, 26, 393-415.

Schmidt, P., S. C. Ahn, and D. Wyhowski (1992): "Comment", *Journal of Business & Economic Statistics*, 10, 10-14.

Sims, C.A. (1972): "Money, Income, and Causality", *American Economic Review*, 62, 540-552.

Wooldridge, J. M. (1995): “Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions,” *Journal of Econometrics*, 68, 115–132.

Wooldridge, J. (1997): “Multiplicative Panel Data Models Without the Strict Exogeneity Assumption”, *Econometric Theory*, 13, 667-678.

Zabel, J. E. (1992): “Estimating Fixed and Random Effects Models with Selectivity,” *Economic Letters*, 40, 269–272.

Zeldes, S.P. (1989): “Consumption and Liquidity Constraints: An Empirical Investigation”, *Journal of Political Economy*, 97, 305-346.

Ziliak, J.P. (1997): “Efficient Estimation with Panel Data when Instruments Are Predetermined: An Empirical Comparison of Moment-Condition Estimators”, *Journal of Business & Economic Statistics*, 15, 419-431.