Journal of Econometrics 163 (2011) 144–162

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Bias corrections for two-step fixed effects panel data estimators

Iván Fernández-Val [a], Francis Vella [b,*]

[a] Boston University, United States

[b] Georgetown University, United States

ARTICLE INFO

Article history:
Received 31 October 2007
Received in revised form
2 March 2011
Accepted 15 March 2011
Available online 24 March 2011

JEL classification:
C23
J31
J51

Keywords:
Panel data
Two-step estimation
Endogenous regressors
Fixed effects
Sample selection bias
Union premium

ABSTRACT

This paper introduces large-T bias-corrected estimators for nonlinear panel data models with both time invariant and time varying heterogeneity. These models include systems of equations with limited dependent variables and unobserved individual effects, and sample selection models with unobserved individual effects. Our two-step approach first estimates the reduced form by fixed effects procedures to obtain estimates of the time varying heterogeneity underlying the endogeneity/selection bias. We then estimate the primary equation by fixed effects including an appropriately constructed control variable from the reduced form estimates as an additional explanatory variable. The fixed effects approach in this second step captures the time invariant heterogeneity while the control variable accounts for the time varying heterogeneity. Since either or both steps might employ nonlinear fixed effects procedures it is necessary to bias adjust the estimates due to the incidental parameters problem. This problem is exacerbated by the two-step nature of the procedure. As these two-step approaches are not covered in the existing literature we derive the appropriate correction thereby extending the use of large-T bias adjustments to an important class of models. Simulation evidence indicates our approach works well in finite samples and an empirical example illustrates the applicability of our estimator.

1. Introduction

The incidental parameters problem arises in the estimation of nonlinear panel models that include unrestricted individual specific effects to control for unobserved time invariant heterogeneity (Neyman and Scott, 1948; Heckman, 1981; Lancaster, 2000; Greene, 2004a). Recent papers, surveyed in Arellano and Hahn (2005) and including Hahn and Kuersteiner (2002, forthcoming), Lancaster (2002), Woutersen (2002), Hahn and Newey (2004), Carro (2006), and Fernández-Val (2009), provide a range of solutions, so-called large-T bias corrections, to reduce the incidental parameters problem in long panels. These papers derive the analytical expression of the bias (up to a certain order of T), which can be employed to adjust the biased fixed effects estimators. Numerical evidence suggests these adjustments eliminate or significantly reduce the bias even in short panels.

While the above papers collectively cover a large class of models, they do not handle endogeneity resulting from unobserved heterogeneity that contains a time varying component. This kind of heterogeneity, which includes time varying endogenous regressors and sample selection, is frequently encountered in empirical investigations. Accordingly we derive large-T bias corrections for panel data models with multiple sources of endogeneity. In particular, we consider a class of models with both time varying and time invariant endogeneity that can be accounted for by including individual effects and a parametric control variable. Specific examples include models with censored endogenous regressors and individual effects, sample selection models with individual effects, and limited dependent variable models with endogenous explanatory variables and individual effects. More generally, our approach covers nonlinear panel data models with predetermined and endogenous regressors, when the endogeneity can be controlled for via individual effects and a parametric control variable.

We provide a computationally simple two-step estimation procedure. We first estimate the reduced form of the time varying heterogeneity underlying the endogeneity/selection bias by fixed effects. We then estimate the primary equation by fixed effects adding an appropriately constructed control variable. Since either or both steps might employ nonlinear fixed effects procedures and the control variable might be a nonlinear function of the individual effects of the reduced form, the incidental parameters problem arises. As the existing bias corrections fail to account for the additional source of incidental parameters bias arising from the fixed effects estimation of the control variable, our main contribution is to extend the large-T bias corrections to systems of equations estimated via two-step fixed effects procedures.

* Corresponding author. Tel.: +1 202 687 5573.
E-mail address: fgv@georgetown.edu (F. Vella).

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.03.002

Below we discuss some papers which have analyzed some of the models we consider here. We differ from these existing studies in our treatment of the time invariant heterogeneity, the assumptions about the properties of the explanatory variables, or the asymptotic framework. Most notable is our treatment of the unobserved individual effects as fixed effects (FE), potentially correlated with the explanatory variables, whereas previous approaches to nonlinear systems of equations in panel data generally assume they are random effects (RE) distributed independently of the explanatory variables. RE estimation by-passes the incidental parameters problem by integrating out the individual effects. This approach, however, has three important shortcomings. First, the independence assumption is not compelling in many applications. In microeconomic studies, for instance, individual effects might capture variations in preferences or technology, and the explanatory variables are often choice variables determined on the basis of this individual heterogeneity. Second, the RE estimators generally require an additional round of integration and this can complicate computation. Finally, the RE procedures require parametric assumptions for the time invariant individual heterogeneity.

A second important feature of our approach is our ability to accommodate weakly exogenous (predetermined) explanators. This extension substantially expands the range of models we can consider by allowing for dynamic feedback effects between the outcomes and the explanatory variables. These effects are not possible under the strict exogeneity assumption which is commonly maintained in the nonlinear panel literature, see, e.g., Rasch (1960), Chamberlain (1980), Manski (1987), Honoré (1992), and Hahn and Newey (2004). Wooldridge (2001), Honoré and Lewbel (2002), Arellano and Carrasco (2003), and Hahn and Kuersteiner (forthcoming) developed one-step estimators for panel data models with predetermined regressors and individual effects. These estimators, however, are not suitable for the models with multiple sources of heterogeneity that we consider.

Note that our procedures are based on large-T asymptotic approximations that are more suitable for moderate or large panels, whereas some of the previous studies remain valid for short panels since they are derived under fixed-T sequences. However, in numerical examples, we find that our bias correction performs well even with only 6 or 8 time periods.

The following section briefly describes some econometric models covered by our approach. Section 3 reviews some existing treatments of bias corrections in non-linear panel data models and extends these corrections to two-step estimators. Section 4 gives the appropriate asymptotic theory. Section 5 provides simulation evidence and Section 6 presents an empirical example. Section 7 adds some concluding remarks. The Appendix contains the proofs of the main results.

2. Panel models with multiple endogeneity

The leading class of econometric models we consider has the following triangular two-equation structure in the observed variables:

$$\begin{aligned} d_{it} &= f_1(x_{1it}, \alpha_{1i}; \theta_1) + \varepsilon_{1it}, \quad \text{(Reduced form equation)} \\ y_{it} &= f_2(d_{it}, x_{2it}, \lambda_{it}, \alpha_{2i}; \theta_2) + \varepsilon_{2it}, \quad \text{(Primary equation)} \end{aligned} \tag{1}$$

for $(i = 1, \dots, n; t = 1, \dots, T)$, where $f_1(\cdot)$ and $f_2(\cdot)$ are known functions up to the finite dimensional parameters θ_1 and θ_2. The endogenous variable of primary interest is y_{it}, and d_{it} is an endogenous explanatory variable or selection indicator. The predetermined explanatory variables are denoted by x_{1it} and x_{2it}; α_{1i} and α_{2i} are unobserved individual effects; λ_{it} is a control variable underlying the endogeneity/selection of d_{it} in the primary equation; and the disturbances are denoted by ε_{1it} and ε_{2it}. An exclusion restriction in x_{2it} relative to x_{1it} ensures that identification does not rely exclusively on nonlinearities of the parametric functions (Cameron and Trivedi, 2005, p. 565). Lags of the observed dependent variables d_{it} and y_{it} may appear in each equation and would be included in x_{1it} and/or x_{2it}.[1] The control variable is assumed to be a known function of the parameters and variables of the reduced form equation, $\lambda_{it} := \lambda(d_{it}, x_{1it}, \alpha_{1i}; \theta_1)$. The form of this function depends on the type of endogeneity/selection and also the nature of the dependent variable in the reduced form. It is usually derived from parametric assumptions about the underlying structural disturbances, typically joint normality, although weaker assumptions can suffice.[2] Below we provide specific examples of econometric models that generate this triangular representation.

Many models in panel data are defined by sequential moment conditions, which correspond to the following restrictions on the disturbances of the system (1):

Assumption 1 (*Sequential Moment Conditions*). The idiosyncratic disturbances ε_{1it} and ε_{2it} satisfy the sequential moment conditions

$$E[\varepsilon_{1it}|x_i(t), \alpha_{1i}] = 0 \quad \text{and} \quad E[\varepsilon_{2it}|d_i(t), x_i(t), \lambda_i(t), \alpha_{2i}] = 0,$$

for $i = 1, \dots, n$; $t = 1, \dots, T$, where $x_i(t) = [x_{1i}(t)', x_{2i}(t)']'$ and $r_i(t) = [r_{i1}, \dots, r_{it}]'$ for $r \in \{d, \lambda, x_1, x_2\}$.

Note that our model is of the FE type because we do not impose any restriction on the joint distribution of α_{1i} and α_{2i} given $x_i(t)$. Assumption 1 indicates that the endogeneity in the primary equation can arise either through the omission of the time invariant unobserved individual effects or through the omission of the time varying control variable. The sequential moment conditions imply that the model is dynamically complete conditional on the individual effects (Wooldridge, 2002, p. 300), and that the explanatory variables are predetermined relative to the disturbances. This is an important departure from the usual strict exogeneity assumption in these models and permits richer dynamic feedbacks from the dependent variable to the explanators. A leading case of predetermined regressors are lags of the dependent variables.

To obtain estimates of the parameters of the model we first estimate the reduced form equation from which we construct the appropriate control variable. We then account for the endogeneity in the primary equation by eliminating the first form, due to the α_2's, through the inclusion of individual fixed effects, and the second, due to the λ's, through the inclusion of the estimated control variable. This approach is computationally more attractive than Full or Partial Maximum Likelihood estimation of the system (1). The computational demands of these models is especially severe due to the presence of fixed effects in both the reduced form and primary equations. Moreover, system estimators, although more efficient, are generally less robust to parametric assumptions than two-step procedures (Wooldridge, 2002, p. 566). The incidental parameters problem may arise in both steps and is further complicated by the inclusion in the second stage of the control variable, which depends on the individual effects of the reduced form equation.

[1] Our framework does not allow for the inclusion of lags of unobserved or latent dependent variables. The inclusion of these variables raises identification issues that are beyond the scope of this paper. See Kyriazidou (2001), Hu (2002) and Gayle and Viauroux (2007) for dynamic sample selection and censored panel models that include lags of latent dependent variables.

[2] There is an extensive literature on the use of control variables to address endogeneity and selection issues in parametric econometric models. In this paper we only derive the control variable for some specific examples, and assume its existence and refer to the literature in which it has been developed in general. See, e.g., Dhrymes (1970), Heckman (1976, 1979), Smith and Blundell (1986), Rivers and Vuong (1988), Blundell and Smith (1989, 1994), and Vella (1993).