Economics Letters 120 (2013) 346–349

Contents lists available at SciVerse ScienceDirect

Economics Letters

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

Avoiding biased versions of Wooldridge's simple solution to the initial conditions problem

Sophia Rabe-Hesketh [a,*], Anders Skrondal [b]

[a] University of California, Berkeley, 3659 Tolman Hall, Berkeley CA 94720-1670, United States
[b] Norwegian Institute of Public Health, P.O. Box 4404, Nydalen, 0403 Oslo, Norway

HIGHLIGHTS

- Wooldridge (2005) proposed a simple solution to the initial conditions problem.
- Popular constrained versions of his auxiliary model can produce severe bias.
- The problem can be avoided by adding initial-period explanatory variables.
- Alternatively, Wooldridge's original model can be used.

ARTICLE INFO

Article history:
Received 31 March 2013
Received in revised form
4 May 2013
Accepted 5 May 2013
Available online 11 May 2013

JEL classification:
C23
C25

Keywords:
Dynamic model
Initial conditions
Panel data
Unbalanced panels

ABSTRACT

Wooldridge (2005) provided a simple and elegant solution to the initial conditions problem for dynamic nonlinear unobserved-effects models. His original auxiliary model includes the time-varying explanatory variables at each period. Unfortunately, a popular constrained version that includes within-means of the explanatory variables can be severely biased. We show that there are several ways to avoid this problem.

1. Introduction

For dynamic probit models with unobserved individual-specific effects, two principal approaches have been proposed for handling the initial conditions problem. Heckman (1981) suggested modeling the initial dependent variable jointly with the subsequent dependent variables, but this approach is not available in standard software. Wooldridge (2005) suggested conditioning on the initial dependent variable by specifying an auxiliary model for the conditional distribution of the unobserved effect given the initial dependent variable and explanatory variables. This approach leads to the standard likelihood for static random-effects probit models with the lagged and initial dependent variables among the regressors.

While Wooldridge's auxiliary model includes values of the time-varying explanatory variables at each period (except the initial period), a more common specification includes the within-means of the time-varying explanatory variables. Reasons for using this approach are rarely given but it may be popular because it is parsimonious and does not require balanced data. However, basing the within-means on all available periods for incomplete panels has not been justified in the literature.

We show that the auxiliary model is overly constrained if it includes within-means of time-varying explanatory variables across all periods, including the initial period (e.g., Jones et al., 2007; Conti and Pudney, 2011; Michaud and Tatsiramos, 2011; Akay, 2012, among many others). The reason is that the conditional distribution of the unobserved effect, given the explanatory variables at all periods (including the initial period), depends more directly on the initial-period explanatory variables than on the explanatory variables at the other periods — in some cases it depends *only* on the initial-period explanatory variables and the initial dependent

* Corresponding author. Tel.: +1 510 642 5287.
E-mail addresses: sophiarh@berkeley.edu (S. Rabe-Hesketh), anders.skrondal@fhi.no (A. Skrondal).

0165-1765/$ – see front matter
http://dx.doi.org/10.1016/j.econlet.2013.05.009

variable. The coefficients of the initial-period explanatory variables should therefore not be constrained to equal the coefficients at the other periods.

Our Monte Carlo experiments show that the constrained model can lead to severe bias for short panels. Similar results for the constrained model led Akay (2012) to conclude incorrectly that "the Wooldridge method can be used instead of Heckman's method only for *moderately long* panels". However, we show that the bias for the constrained model practically disappears when the initial-period explanatory variables are included as additional regressors or when Wooldridge's original auxiliary model is used.

The paper proceeds as follows. Section 2 introduces the model assumptions, describes different versions of the Wooldridge method, and explains why the constrained auxiliary model is poorly specified. Monte Carlo experiments are performed in Section 3 to compare the constrained model with justified alternatives.

2. Specifying the auxiliary model

A dynamic probit unobserved-effects model for the dependent variable y_{it} for individual i ($i = 1, \ldots, N$) at time period t can be written as

$$y_{it}^* = \mathbf{z}_{it}'\boldsymbol{\gamma} + \rho y_{i,t-1} + c_i + u_{it}, \quad u_{it} \sim \mathcal{N}(0,1),$$
$$y_{it} = 1(y_{it}^* > 0). \tag{1}$$

Following Wooldridge (2005), it is assumed that the time-varying explanatory variables $\mathbf{z}_{it}$ are strictly exogenous, conditional on the individual-specific unobserved effect c_i. We assume that c_i is normally distributed and sometimes that it is independent of all $\mathbf{z}_{it}$. The process starts at period $s < 1$ and is observed at periods $t = 1, \ldots, T$.[1]

Wooldridge (2005) addresses the initial conditions problem by modeling y_{it} at periods $t = 2, \ldots, T$, given the initial dependent variable y_{i1} and explanatory variables. He specifies the required conditional density of the unobserved effect c_i via the auxiliary model

$$\mathbf{W}: \; c_i = \alpha_0 + \alpha_1 y_{i1} + \mathbf{z}_i^{+\prime}\boldsymbol{\alpha}_2 + a_i, \tag{2}$$

where $\mathbf{z}_i^+ = (\mathbf{z}_{i2}', \ldots, \mathbf{z}_{iT}')'$. Here and henceforth, a_i is normal with mean 0 and variance σ_a^2, given the regressors in the auxiliary model.

The constrained model (e.g., Akay, 2012)

$$\mathbf{C}: \; c_i = \alpha_0 + \alpha_1 y_{i1} + \bar{\mathbf{z}}_i'\boldsymbol{\alpha}_2 + a_i \tag{3}$$

uses within-means based on all periods including the first, $\bar{\mathbf{z}}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_{it}$, instead of $\mathbf{z}_i^+$. Wooldridge (2005) showed that consistency requires correct specification of the conditional distribution of c_i given y_{i1} and $\mathbf{z}_i^{+\prime}$, whereas the constrained model also conditions on $\mathbf{z}_{i1}$.

We now show why the constrained model is poorly specified. The correct conditional distribution $f(c_i|y_{i1}, \mathbf{z}_i)$ of c_i given y_{i1} and $\mathbf{z}_i = (\mathbf{z}_{i1}', \mathbf{z}_i^{+\prime})'$, implied by the model, is

$$f(c_i|y_{i1}, \mathbf{z}_i) \propto \mathrm{P}(y_{i1}|\mathbf{z}_i, c_i) f(c_i|\mathbf{z}_i) \tag{4}$$
$$= \mathrm{P}(y_{i1}|\mathbf{z}_i, c_i) f(c_i) \quad \text{if } c_i \perp \mathbf{z}_i. \tag{5}$$

Here the required conditional distribution of y_{i1} is

$$\mathrm{P}(y_{i1}|\mathbf{z}_i, c_i) = \sum_{y_{i0}\in\{0,1\}} \mathrm{P}(y_{i1}|\mathbf{z}_{i1}, y_{i0}, c_i)\mathrm{P}(y_{i0}|\mathbf{z}_i, c_i), \tag{6}$$

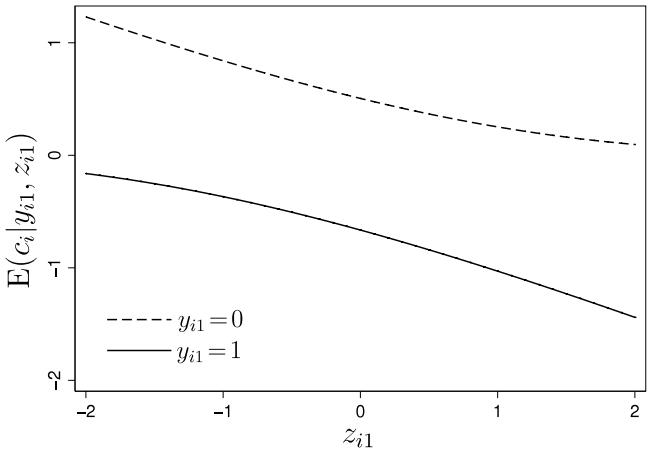

Fig. 1. Relationship between $\mathrm{E}(c_i|y_{i1}, z_{i1})$ and z_{i1} for $y_{i1} = 0, 1$ ($c_i \sim \mathcal{N}(0,1)$, $s = 0$, $\mathrm{P}(y_{i0} = 1|\mathbf{z}_i, c_i) = 0.5$, $\gamma_0 = 0$, $\gamma_1 = 1$, $\rho = 0.5$).

where $\mathrm{P}(y_{i1}|\mathbf{z}_{i1}, y_{i0}, c_i)$ follows from (1) and y_{i0} is the last presample dependent variable.

Notice that $\mathrm{P}(y_{i1}|\mathbf{z}_i, c_i)$ and hence $f(c_i|y_{i1}, \mathbf{z}_i)$ depends directly on $\mathbf{z}_{i1}$ via $\mathrm{P}(y_{i1}|\mathbf{z}_{i1}, y_{i0}, c_i)$, after allowing for the dependence on $\mathbf{z}_i$ via $\mathrm{P}(y_{i0}|\mathbf{z}_i, c_i)$. In fact, $f(c_i|y_{i1}, \mathbf{z}_i)$ depends only on y_{i1} and $\mathbf{z}_{i1}$ (and not on $\mathbf{z}_{it}$, $t > 1$) if c_i is independent of $\mathbf{z}_i$ and either the $\mathbf{z}_{it}$ are independent over time[2] or the process starts at $s = 0$. In these cases, it follows from (5) and (6) that

$$\mathrm{E}(c_i|y_{i1}, \mathbf{z}_i) = \frac{\int c_i f(c_i)\left[\sum_{y_{i0}\in\{0,1\}} \mathrm{P}(y_{i1}|\mathbf{z}_{i1}, y_{i0}, c_i)\mathrm{P}(y_{i0}|c_i)\right] \mathrm{d}c_i}{\int f(c_i)\left[\sum_{y_{i0}\in\{0,1\}} \mathrm{P}(y_{i1}|\mathbf{z}_{i1}, y_{i0}, c_i)\mathrm{P}(y_{i0}|c_i)\right] \mathrm{d}c_i}. \tag{7}$$

Fig. 1 shows, for univariate z_{it}, that this conditional expectation[3] can depend strongly on z_{i1} and y_{i1}.

We hypothesize that the constrained model C performs poorly because it implicitly sets the coefficients of the initial explanatory variables equal to the coefficients for the subsequent periods, which is at odds with the form of the correct distribution. Making minimal changes to C, the proposed model P relaxes the unrealistic constraint by including the initial-period explanatory variables $\mathbf{z}_{i1}$ as additional regressors:

$$\mathbf{P}: \; c_i = \alpha_0 + \alpha_1 y_{i1} + \bar{\mathbf{z}}_i'\boldsymbol{\alpha}_2 + \mathbf{z}_{i1}'\boldsymbol{\alpha}_3 + a_i. \tag{8}$$

Here $\bar{\mathbf{z}}_i$ could be replaced by the mean $\bar{\mathbf{z}}_i^+ = \frac{1}{T-1}\sum_{t=2}^{T}\mathbf{z}_{it}$ that does not include the initial-period explanatory variables.[4] Alternatively, we can easily relax the constraint by omitting the initial-period explanatory variables from the within-means

$$\mathbf{Q}: \; c_i = \alpha_0 + \alpha_1 y_{i1} + \bar{\mathbf{z}}_i^{+\prime}\boldsymbol{\alpha}_2 + a_i. \tag{9}$$

Such a model is sometimes used (e.g., Drakos and Konstantinou, 2013), but it is often not clear whether the within-means are defined as $\bar{\mathbf{z}}_i^+$ or $\bar{\mathbf{z}}_i$.

[1] This notation differs from Wooldridge's where the first observed period is $t = 0$.

[2] For proof and simulation evidence, see Supplementary Materials.

[3] Integrals evaluated by 50-point adaptive quadrature (Rabe-Hesketh et al., 2005)

[4] If $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ are the coefficients when $\bar{\mathbf{z}}_i$ is used, then $\boldsymbol{\alpha}_2^* = (T-1)\boldsymbol{\alpha}_2/T$ and $\boldsymbol{\alpha}_3^* = \boldsymbol{\alpha}_3 + \boldsymbol{\alpha}_2/T$ are the corresponding coefficients when $\bar{\mathbf{z}}_i^+$ is used.

Table 1
Results of Monte Carlo experiments for models C, P, and Q.

T	Relative bias (%)			RMSE			Relative bias (%)			RMSE		
	C	P	Q	C	P	Q	C	P	Q	C	P	Q
Exogenous $\mathbf{z}_i$												
	MCE 1: $z_{it} \sim$ i.i.d. $\mathcal{N}(0,1)$						MCE 2: $z_{it} \sim$ i.i.d. $(\chi^2(1)-1)/\sqrt{2}$					
3	39.225	3.219	3.337	0.381	0.353	0.356	25.598	1.577	1.902	0.376	0.369	0.371
4	14.101	−2.976	−3.155	0.224	0.211	0.212	9.540	−1.817	−1.713	0.220	0.209	0.210
5	12.476	2.384	2.276	0.181	0.167	0.167	3.176	−3.286	−3.194	0.172	0.170	0.170
8	1.627	−1.788	−1.790	0.115	0.114	0.114	1.767	−0.380	−0.389	0.108	0.107	0.107
20	0.157	−0.350	−0.358	0.058	0.058	0.058	0.476	0.147	0.164	0.058	0.058	0.058
	MCE 3: AR(1) process for z_{it}						MCE 4: Nerlove process for z_{it}					
3	17.104	−5.529	−17.170	0.380	0.384	0.402	−2.542	−4.048	−6.343	0.417	0.417	0.420
4	8.770	−0.526	−5.925	0.241	0.233	0.235	−0.360	−0.740	−1.895	0.241	0.241	0.241
5	7.497	1.579	−1.133	0.175	0.171	0.169	0.495	0.436	−0.176	0.185	0.185	0.184
8	1.170	−0.898	−1.574	0.110	0.109	0.109	1.628	1.692	1.530	0.112	0.112	0.112
20	0.438	0.114	0.071	0.064	0.064	0.064	0.177	0.195	0.186	0.084	0.084	0.084
Endogenous $\mathbf{z}_i$												
	MCE 1: $z_{it} \sim$ i.i.d. $\mathcal{N}(0,1)$						MCE 2: $z_{it} \sim$ i.i.d. $(\chi^2(1)-1)/\sqrt{2}$					
3	36.136	0.615	0.708	0.392	0.396	0.393	22.712	1.194	1.243	0.357	0.354	0.355
4	12.125	−4.542	−4.773	0.224	0.212	0.213	8.711	−1.268	−1.253	0.229	0.221	0.222
5	12.123	2.094	1.977	0.188	0.174	0.174	3.189	−2.664	−2.625	0.178	0.176	0.176
8	2.309	−1.089	−1.082	0.110	0.109	0.109	2.379	0.382	0.340	0.115	0.113	0.113
20	0.157	−0.343	−0.354	0.059	0.059	0.059	0.189	−0.127	−0.117	0.059	0.059	0.059
	MCE 3: AR(1) process for z_{it}						MCE 4: Nerlove process for z_{it}					
3	15.056	−6.279	−12.217	0.389	0.397	0.405	−3.893	−5.284	−8.346	0.419	0.420	0.425
4	8.627	−0.539	−3.836	0.243	0.236	0.236	−0.399	−0.793	−2.235	0.252	0.252	0.251
5	5.080	−0.638	−2.565	0.186	0.183	0.182	−0.634	−0.709	−1.403	0.179	0.179	0.179
8	1.923	−0.103	−0.705	0.116	0.115	0.115	1.523	1.562	1.401	0.115	0.115	0.114
20	0.585	0.263	0.219	0.063	0.063	0.063	−0.670	−0.663	−0.668	0.069	0.068	0.069

Note: Underlined estimates differ significantly from 0 using one-sample t tests at the 5% level.

3. Monte Carlo experiments

To confirm the poor performance of model C found by Akay (2012) and investigate whether it is due to the constraint that the coefficients of the initial explanatory variables equal the coefficients for the subsequent periods, we perform Monte Carlo experiments for models C, P, and Q. Although Wooldridge (2005) showed that consistency does not require conditioning on $\mathbf{z}_{i1}$, we are interested in comparing P and Q because the distribution of c_i given y_{i1} and $\mathbf{z}_i$ depends more directly on $\mathbf{z}_{i1}$ than on $\mathbf{z}_i^+$. If $\mathbf{z}_{i1}$ is strongly predictive of c_i, including $\mathbf{z}_{i1}$ may reduce bias and/or increase efficiency. We also investigate whether the equality constraints between the coefficients of the explanatory variables at periods 2 to T should be relaxed. Relaxing these constraints for model Q gives Wooldridge's auxiliary model W in (2) and relaxing them for model P gives the augmented Wooldridge model W^* that includes $\mathbf{z}_{i1}$ as additional regressors.

In all experiments, the process starts at $s = -24$, with $y_{is} \sim$ Bernoulli(0.5). The subsequent responses follow model (1) with one time-varying explanatory variable z_{it}. We consider both exogenous and endogenous $\mathbf{z}_i$, with $c_i \sim \mathcal{N}(0, \sigma_c^2)$ and $c_i \sim \mathcal{N}(0.5\bar{z}_i, \sigma_c^2)$, respectively. The parameter values are $\gamma_0 = 0$, $\gamma_1 = 1$, $\rho = 0.5$, and $\sigma_c = 1$ for experiments 1, 2, and 3 and $\gamma_0 = 4$, $\gamma_1 = -1$, $\rho = 0.5$, and $\sigma_c = 1$ for experiment 4. The sample data are observed from time $t = 1$ for $T = 3, 4, 5, 8$, and 20 periods and for $N = 200$ individuals. The Monte Carlo experiments differ in the processes that generate the time-varying explanatory variable z_{it}:

MCE1 (Independent normal): $z_{it} \sim$ i.i.d. $\mathcal{N}(0,1)$
MCE2 (Independent skewed): $z_{it} \sim$ i.i.d. $(\chi^2(1)-1)/\sqrt{2}$
MCE3 (AR(1) process): $z_{it} = 0.5z_{i,t-1} + \psi_{it}$, $\psi_{it} \sim \mathcal{N}(0,1)$, and $z_{is} = \psi_{is}$
MCE4 (Nerlove process): $z_{it} = 0.1t + 0.5z_{i,t-1} + \psi_{it}$, $\psi_{it} \sim U[-1/2, 1/2]$, and $z_{is} \sim U[-3, 2]$.

For each of the 40 simulation conditions (exogenous versus endogenous $\mathbf{z}_i$, four processes for $\mathbf{z}_i$, and five values of T), we simulate 500 datasets, each with new values of $\mathbf{z}_i$ and estimate the five models using `xtprobit` in Stata with 20-point adaptive quadrature.

The estimated relative bias (in percent) and root mean squared error (RMSE) for ρ are given in Table 1 for models C, P, and Q. For the exogenous case and model C, the simulations are replications of those reported by Akay (2012) and the estimated bias never differs significantly from his at the 5% significance level (using two-sample t tests). Our results confirm that model C produces upward biased estimates of ρ, particularly when the z_{it} are independent over time (MCE 1 and 2). The relative bias is significant at the 5% level (using one-sample t tests) for MCE 1, 2, and 3 for panels of length up to $T = 5$ and substantial for panels of length $T = 3$ and $T = 4$. Model P generally produces small estimated relative bias which does not differ significantly from 0 except for two of the 40 simulation conditions. The relative bias is comparable to that found by Akay (2012) for the Heckman (1981) method in the exogenous case. However, the RMSEs are larger than for Heckman's approach, particularly for short panels, except for the Nerlove process. Model Q performs similarly to model P except for the AR(1) process with short panels where it has greater bias. As shown in Table 2, Wooldridge's model W performs well, as was also found by Arulampalam and Stewart (2009), with estimated relative bias and RMSE similar to model P. The augmented model W* does not appear to perform better than model W.

4. Conclusion

A popular constrained version of Wooldridge's (2005) simple solution to the initial conditions problem, which uses within-means of time-varying explanatory variables, performs poorly for short panels if the means are based on all periods, including the initial period. This problem can be avoided by either including the initial-period explanatory variables as additional regressors or by using Wooldridge's original auxiliary model.

Table 2
Results of Monte Carlo experiments for models W and W*.

T	Relative bias (%)		RMSE		Relative bias (%)		RMSE	
	W	W*	W	W*	W	W*	W	W*
Exogenous $\mathbf{z}_i$								
	MCE 1				MCE 2			
3	1.969	1.481	0.374	0.370	2.128	1.707	0.389	0.388
4	−2.853	−2.680	0.219	0.218	−1.562	−1.637	0.215	0.215
5	2.577	2.754	0.169	0.169	−2.889	−2.979	0.170	0.169
8	−1.683	−1.682	0.115	0.114	−0.054	−0.035	0.107	0.107
20	−0.281	−0.276	0.058	0.058	0.310	0.292	0.058	0.058
	MCE 3				MCE 4			
3	−7.020	−7.735	0.403	0.400	−4.052	−4.171	0.417	0.415
4	−0.313	−0.215	0.238	0.238	−0.717	−0.699	0.241	0.241
5	1.679	1.416	0.175	0.175	0.368	0.404	0.185	0.185
8	−0.967	−0.955	0.110	0.110	1.624	1.642	0.112	0.112
20	0.169	0.184	0.064	0.064	0.255	0.224	0.084	0.084
Endogenous $\mathbf{z}_i$								
	MCE 1				MCE 2			
3	−1.089	−1.270	0.424	0.422	0.669	0.547	0.366	0.365
4	−4.599	−4.332	0.215	0.215	−1.018	−0.999	0.227	0.227
5	2.358	2.528	0.177	0.177	−2.432	−2.472	0.176	0.176
8	−0.963	−0.967	0.109	0.109	0.559	0.611	0.113	0.113
20	−0.270	−0.261	0.059	0.059	0.014	0.005	0.059	0.059
	MCE 3				MCE 4			
3	−8.838	−8.989	0.415	0.413	−4.949	−5.132	0.421	0.417
4	−0.320	−0.220	0.243	0.242	−0.699	−0.620	0.251	0.251
5	−0.805	−0.969	0.185	0.185	−0.781	−0.745	0.180	0.180
8	−0.204	−0.161	0.115	0.115	1.535	1.553	0.115	0.115
20	0.320	0.333	0.062	0.063	−0.631	−0.640	0.069	0.069

Note: Underlined estimates differ significantly from 0 using one-sample t tests at the 5% level.

Appendix. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.econlet.2013.05.009.

References

Akay, A., 2012. Finite-sample comparison of alternative methods for estimating dynamic panel data models. Journal of Applied Econometrics 27 (7), 1189–1204.

Arulampalam, W., Stewart, M.B., 2009. Simplified implementation of the Heckman estimator of the dynamic probit model and a comparison with alternative estimators. Oxford Bulletin of Economics and Statistics 71 (5), 659–681.

Conti, G., Pudney, S., 2011. Survey design and the analysis of satisfaction. The Review of Economics and Statistics 93 (3), 1087–1093.

Drakos, K., Konstantinou, P.T., 2013. Investment decisions in manufacturing: assessing the effects of real oil prices and their uncertainty. Journal of Applied Econometrics 28 (1), 151–165.

Heckman, J.J., 1981. The incidental parameters problem and the problem of initial conditions in estimating a discrete time — discrete data stochastic process. In: Manski, C.F., McFadden, D.L. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA, pp. 179–195.

Jones, A.M., Rice, N., Bago d'Uva, T., Balia, S., 2007. Applied Health Economics. Routledge, London.

Michaud, P.-C., Tatsiramos, K., 2011. Fertility and female employment dynamics in Europe: the effect of using alternative econometric modeling assumptions. Journal of Applied Econometrics 26 (4), 641–668.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., 2005. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. Journal of Econometrics 128 (2), 301–323.

Wooldridge, J.M., 2005. Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. Journal of Applied Econometrics 20 (1), 39–54.