Universitat Pompeu Fabra Barcelona
Department of Economics and Business

Economics Working Paper Series

Working Paper No. 1631

Simple methods for consistent estimation of dynamic panel data sample selection models

Majid al Sadoon, Sergi Jiménez-Martín,
and José M. Labeaga

January 2019

Simple methods for consistent estimation of dynamic panel data sample selection models *

Majid al Sadoon† Sergi Jiménez-Martín‡ José M. Labeaga§

January 2019

Abstract

We analyse the properties of generalised method of moments-instrumental variables (GMM-IV) estimators of AR(1) dynamic panel data sample selection models. We show the consistency of the first-differenced GMM-IV estimator uncorrected for sample selection of Arellano and Bond (1991) (a property also shared by the Anderson and Hsiao,1982, proposal). Alternatively, the system GMM-IV estimator (Arellano and Bover, 1995, and Blundell and Bond, 1998) shows a moderate bias. We perform a Monte Carlo study to evaluate the finite sample properties of the proposed estimators. Our results confirm the absence of bias of the Arellano and Bond estimator under a variety of circumstances, as well as the small bias of the system estimator, mostly due to the correlation between the individual heterogeneity components in both the outcome and selection equations. However, we must not discard the system estimator because, in small samples, its performance is similar to or even better than that of the Arellano-Bond. These results hold in dynamic models with exogenous, predetermined or endogenous covariates. They are especially relevant for practitioners using unbalanced panels when either there is selection of unknown form or when selection is difficult to model.

JEL Codes: J52, C23, C24

Keywords: Panel data, sample selection, dynamic model, generalized method of moments

1 Introduction

The problems of self-selection, non-response and attrition are common in datasets containing economic variables. Although dealing with them in cross-sections results in manageable models, correlated heterogeneity together with endogenous attrition or selection complicates the models with

*We are grateful to the Spanish Ministry of Economy for financial support through projects ECO2014-52238-R, and ECO2017-83668-R. We are also grateful to Manuel Arellano, Badi Baltagi and María Rochina-Barrachina for some very useful comments, to the seminar audience at UPF and to participants at the 2015 Annual Meeting of the International Applied Econometric Society held at Thessaloniki, especially to Frank Windmeijer and Juan M. Rodríguez-Poo. All remaining errors are our responsibility. The usual disclaimer applies.

†Durham University Bussiness School

‡Universitat Pompeu Fabra and BGSE. Corresponding author: sergi.jimenez@upf.edu

§UNED

unbalanced panel data (Baltagi, 2005). Many studies have dealt with unobserved heterogeneity and selectivity simultaneously, as we will review in the next section. The increasing availability of large longitudinal datasets and the development of new methods make these approaches likely to be used more frequently in the future. In this context, we believe that it is important to highlight advantages and problems in the performance of different estimators and to draw researchers' attention to potential pitfalls in using them in empirical studies.

In this paper we focus on the estimation of the AR(1) dynamic panel data sample selection model, when the selection process is either static or dynamic. Note, however, that all the results nicely extend to the model with covariates. We assume a typical model for the outcome of interest and consider different assumptions for the selection equation. The error components of both equations can be correlated with a very general correlation structure. Departing from the simplest situation, we present an exercise including all important features in the model one by one to test its individual and joint effects on the bias of generalised method of moments (GMM) estimators. Thus, our exercise can be viewed as a guide for applied researchers on the cost (in terms of bias) of estimating the model on an unbalanced selected panel or on a balanced panel constructed by not considering sample selection.

We show that the uncorrected (for selection) generalized method of moments instrumental variables (GMM-IV) Arellano and Bond (AB, 1991), as well as the less efficient Anderson and Hsiao (AH, 1982), estimators are consistent regardless of whether selection is exogenous or endogenous. Furthermore, we show that the additional orthogonality restrictions implied by the system GMM estimator (Arellano and Bover, 1995; Blundell and Bond, 1998) are not valid under endogenous selection. However, the inconsistency of the estimator is very small and hardly induces bias in the estimator, even and especially in small samples, when the time-invariant heterogeneity components in the outcome and selection equations are not correlated. All these results extend to the model with exogenous, predetermined or endogenous covariates.

These estimators are then evaluated using Monte Carlo methods relaxing or imposing a variety of assumptions. All of our results suggest the non-necessity of correcting the first difference AB estimates in the selected sample. For those that still pursue full elimination of the small bias of the system estimator, we evaluate simple corrections for selectivity in the equations in levels, which are based on estimates of the correlation between the heterogeneity components of corrections resulting in typical binomial probit models adjusted for each cross-section. The corrected outcome equation is then adjusted by a system GMM estimator that can be implemented using standard software, although, of course, it requires computing the correct standard errors.[1]

Thus, our exercise provides a general picture implying little necessity to correct for selectivity when we allow for moderate or even high degrees of selection and the selection equation is either static or dynamic. Our results also apply to outcome equations with exogenous, predetermined or endogenous covariates. We submit the model in the Monte Carlo exercise to several sensitivity

[1]For instance, following Terza (2016).

checks by relaxing various maintained assumptions and show that they are very robust except in the case in which the ratio of variances of the heterogenous component to the idiosyncratic error is high. Overall, we believe that these results could be especially relevant for practitioners in cases involving sample selection of unknown form, when the selection process is difficult to model or when exclusion restrictions are not available.

The rest of the paper is organised as follows. Section 2 provides a review of the literature and presents the general model and the estimation methods. The performance of the proposed estimators is tested in Section 3 in which we present a Monte Carlo study of the finite sample average bias of GMM-IV estimators as well as a sensitivity analysis of some assumptions. In Section 4, we present and evaluate simple alternatives for correction. In Section 5, we compare the different estimators in an empirical exercise, which uses the same data of Semykina and Wooldridge (2013, SW) or Lai and Tsai (2016). Finally, Section 6 concludes.

2 Modelling strategy

2.1 Previous literature

The problem of endogenous selection is common in the empirical economic literature using panel data and it has also received attention in theoretical econometrics models. Starting with Verbeek and Nijman (1992), who proposed tests of selection bias either with or without allowance for correlation between the unobserved effects and explanatory variables, a number of proposals considering unobserved heterogeneity and selectivity simultaneously have appeared. Some of them, such as Wooldridge (1995) and Rochina-Barrachina (1999), proposed new methods for estimating the sample selection model with correction under strict exogeneity. Kyriazidou (1997) corrected for selection bias using a semiparametric approach based on a conditional exchangeability assumption and Lai and Tsay (2016) proposed maximum simulated likelihood methods. On the other hand, Vella and Verbeek (1998), Charlier *et al.* (2001) and Semykina and Wooldridge (2010) allowed for endogenous explanatory variables. Finally, Semykina and Wooldridge (2018) proposed estimation procedures for discrete choice panel data models.[2]

Dynamics appeared for the first time in the work of Kyriazidou (2001), who extended her previous proposal. More recently, Semykina and Wooldridge (2013) proposed new two-stage random effects strategies for estimating panel data models in the presence of endogeneity, dynamics and selection.[3] Note, however, that the validity of Semykina and Wooldridge's method is based on the validity of the assumption of correlation of the heterogeneity components and the initial condition. Because none of the previous papers suggested a preferred, simplified, or dominant method, our aim here is to provide solutions easily applicable from the point of view of applied practitioners.

[2]In another strand of research, theoretical papers have explored bias-corrected estimators for the static case (Fernández-Val and Vella, 2007).

[3]In the dynamic case, semiparametric alternatives were studied by Gayle and Viauroux (2007) and Sasaki (2015), while maximum likelihood methods were explored by Raymond *et al.* (2010).

The various methods have been applied to a number of empirical studies. Charlier *et al.* (2001) studied housing expenditure by households. Jones and Labeaga (2004) selected a sample of non-smokers using the variable addition test of Wooldridge (1995) and then estimated Tobit-type models on the sample of smokers and potential smokers using GMM and minimum distance (MD) methods. González-Chapela (2007) used GMM to estimate the effect of recreational goods on male labour supply. Winder (2004) used instrumental variables to account for endogeneity of some regressors in earnings equations for females. Jiménez-Martín (2006) estimated dynamic wage equations and tested the possibility of differences between strikers and non-strikers. Dustmann and Rochina-Barrachina (2007) estimated females' wage equations extending Rochina-Barrachina (1999). More recently, Semykina and Wooldridge (2010, 2013) applied their methods to estimate earnings equations for females. Finally, Semykina and Wooldridge (2018) applied discrete choice sample selection panel data models to the analysis of pension coverage among white females in the US.

Because it is likely these approaches will be used more frequently in the future, we believe that it is important to highlight properties, advantages and problems of the various methods, as well as their pitfalls and performance in applied studies. This is precisely what we aim to do in this paper. First, we show the consistency of the AB GMM-IV estimator (and implicitly of the AH estimator) and establish a bound for the system GMM-IV in the worse-case scenario of endogenous selection. Then, we carry out a Monte Carlo exercise to examine the performance of each method under alternative assumptions. Finally, we compare the different estimates in an empirical exercise.

2.2 The model

Consider the following AR(1) panel data model with unobserved heterogeneity:

$$y_{it}^{*} = \rho y_{it-1} + \alpha_i + \varepsilon_{it} \tag{1}$$

for $i = 1, ..., N$ and $t = 1, ..., T$. α_i is an individual heterogeneous component independent of the idiosyncratic error ε_{it}, and ρ is a parameter to estimate. In the case of selection, the variable of interest is partially observed, and it is normal to specify an observability or selection rule of the form:

$$d_{it}^{*} = z_{it}\gamma + \eta_i + u_{it} \tag{2}$$

where η_i is a term capturing unobserved individual heterogeneity, z_{it} (which also includes a constant) is a vector of strictly exogenous regressors once we allow them to be correlated with η_i, and u_{it} is an error term. The observed indicator d_{it} is:

$$d_{it} = 1[d_{it}^{*} > 0] = 1[z_{it}\gamma + \eta_i + u_{it} > 0] \tag{3}$$

in a way such that $d_{it} = 1$ if y_{it}^{*} is observed and zero otherwise. The selection equation (2) could

also contain a lagged observed indicator (d_{it-1}), which we ignore for the moment to keep notation as simple as possible.

The error components in equation (2) are related to the error components in the selection equation as follows:

$$\alpha_i = \alpha_i^0 + \theta_0 \eta_i \tag{4}$$

$$\varepsilon_{it} = \varepsilon_{it}^0 + \vartheta_0 u_{it} \tag{5}$$

where α_i^0 and ε_{it}^0 are assumed to be normally distributed and θ_0 and ϑ_0 are the parameters introducing correlation. In the case that they are both zero, there is exogenous sample selection (and, thus, likely unbalancedness in the estimation sample). Alternatively, when any of them is different from zero, there is endogenous sample selection.

It is well known that in the absence of endogenous selection and for the typical situation of N large and T small, model (1) in first differences is usually estimated by IV, as firstly introduced by Anderson and Hsiao (1982). Arellano and Bond (1991), among others, proposed a more efficient GMM-IV estimator, while Arellano and Bover (1995) extended the previous GMM-IV approach to include equations in levels and proposed the estimation of the whole model using system GMM. As noted by Blundell and Bond (1998) in the case of an AR(1) with highly persistent time series correlation, first-differencing could lead to a weak instruments problem (see Roodman, 2009). Then, the use of equations in levels could become important to improve efficiency.

2.3 Estimation of the outcome equation under endogenous selection

In the presence of endogenous sample selection, researchers are tempted to think that the way to proceed is analogous to the method used for the standard static case (as described by Wooldridge, 1995, and others). First, to correct the problem of endogenous selection induced by the correlation of the errors in both equations, and then, to estimate the model.

The dynamics of the model and its transformation to first differences imply that the sample is conditional to observing the outcome for at least three consecutive periods and the amount of data lost depends on the degree of selection. If we use the system GMM estimator, the estimating sample differs by equation. For the equations in levels, we have:

$$y_{it}^* = \rho y_{it-1} + \alpha_i + \varepsilon_{it} \quad if \quad d_{it}, d_{it-1} = 1 \tag{6}$$

for samples of two consecutive periods. For the first-differenced equations, we have:

$$\Delta y_{it}^* = \rho \Delta y_{it-1} + \Delta \varepsilon_{it} \quad if \quad d_{it}, d_{it-1}, d_{it-2} = 1 \tag{7}$$

and we keep for estimation only individuals observed over three consecutive periods.

Note that in those cases in which there is no endogenous selection, because GMM-IV methods are based on instruments that are uncorrelated with both the errors in levels ε_{it} and in first-differences $\Delta\varepsilon_{it}$, it should be feasible to recover consistent estimates of the model parameters (see Arellano and Bond, 1991). For the first-differenced equations all the values of y lagged at least twice are valid instruments because $E(\Delta\varepsilon_{it}y_{it-k}) = 0, k = 2, 3, ...$ Note that in order to have a valid instrument for the first- differenced equations, we do not need to impose any further restriction on the data (ie. we do not need to reserve extra periods of data to construct the instrument). For the levels equations, Δy_{it-1} is also valid, because $E((\alpha_i + \varepsilon_{it})\Delta y_{it-1}) = 0$. However, in this case, we have to add an additional restriction to the dataset, thereby equalizing the sample condition with the one previously required.

Under endogenous sample selection, the validity of the above instruments is questionable because, for the first-differenced errors and the first lagged instrument available, the following orthogonality conditions have to hold:

$$E(\Delta\varepsilon_{it}y_{it-2}/z_{it}, d_{it} = d_{it-1} = d_{it-2} = 1) = 0 \tag{8}$$

which is stronger than the sample condition imposed in the standard case. Note that when this restriction holds, it also holds for $t-3$ and backward lags. For the equations in levels, we need the following:

$$E((\alpha_i + \varepsilon_{it})\Delta y_{it-1}/z_{it}, d_{it} = d_{it-1} = d_{it-2} = 1) = 0 \tag{9}$$

which is also stronger than in the general case.

Our initial guess, based on previous work by Arellano *et al.* (1999), is that because the final estimating sample is selected on positives for at least three consecutive previous periods, the need to correct is greatly reduced.[4] However, in the next section, we show that condition (8) holds, so both the standard IV estimator of AH and the GMM-IV first-differenced estimator of AB are consistent even under endogenous selection. Alternatively, we show that condition (9) does not hold, so the GMM-IV system estimator has a source of inconsistency. However, we will show that, under very general circumstances, this inconsistency and the consequent bias are small. In this context, we show that in those cases in which the AB estimator does not work well (small N, large autoregressive coefficient), the system estimator is highly recommended. Note that these result also imply that we only need to correct the equations in levels (and not the first-differenced ones) in those cases in which we are interested in obtaining a truly consistent version of the system estimator in the presence of sample selection.

[4] Arellano *et al.* (1999) proposed the estimation of sample selection models conditioning on exogenous positive past outcomes and showed that the degree of selection is significantly reduced in economic models with persistence.

2.4 Consistency of the GMM estimators under endogenous sample selection

2.4.1 The pure autoregressive model

Let us start with a minor modification of the AR(1) model presented in equations (1) and (2) to be more precise with the assumptions:

$$y_{it}^* = \alpha_i + \rho_0 y_{it-1}^* + \varepsilon_{it} \tag{10}$$
$$d_{it} = 1(\eta_i + \gamma_0 z_{it} + u_{it} > 0) \tag{11}$$
$$\alpha_i = \alpha_i^0 + \theta_0 \eta_i \tag{12}$$
$$\varepsilon_{it} = \varepsilon_{it}^0 + \vartheta_0 u_{it} \tag{13}$$

The exogenous random variables z_{it}, α_i^0, ε_{it}^0, η_i, and u_{it} are assumed to be i.i.d. and independent of each other with finite second moments. We assume that $|\rho| < 1$ and y_{it}^* is the stationary causal solution to the AR(1) model, $y_{it}^* = \frac{\alpha_i}{1-\rho_0} + \sum_{j=0}^{\infty} \rho_0^j \varepsilon_{it-j}$. We also assume that $E(\varepsilon_{it}^0) = E(u_{it}) = 0$. The observed data is the set of y_{it}^* for which $d_{it} = 1$.

Let $\Delta\varepsilon_{it}(\rho) = \Delta y_{it}^* - \rho \Delta y_{it-1}^*$. The natural moment conditions to consider would be $E(y_{is}^* \Delta\varepsilon_{it}(\rho)) = 0$ for $s + 2 \leq t$ iff $\rho = \rho_0$. However, because y_{it}^* is not always observed, the moment cannot be estimated. The next best option is to try to show $E(s_{ist} y_{is}^* \Delta\varepsilon_{it}(\rho)) = 0$ iff $\rho = \rho_0$, where s_{ist} is defined as

$$s_{ist} = d_{it} d_{it-1} d_{it-2} d_{is} \tag{14}$$

Thus, $s_{ist} = 1$ if and only if all y_{is}^* and $\Delta\varepsilon_{it}(\rho)$ are observed. Now, write

$$E(s_{ist} y_{is}^* \Delta\varepsilon_{it}(\rho)) = E(s_{ist} y_{is}^* (\Delta y_{it}^* - \rho \Delta y_{it-1}^*)) \tag{15}$$
$$= E(s_{ist} y_{is}^* (\rho_0 \Delta y_{it-1}^* + \Delta\varepsilon_{it} - \rho \Delta y_{it-1}^*)) \tag{16}$$
$$= (\rho_0 - \rho) E(s_{ist} y_{is}^* \Delta y_{it-1}^*) + E(s_{ist} y_{is}^* \Delta\varepsilon_{it}) \tag{17}$$

Identification requires that $E(s_{ist} y_{is}^* \Delta y_{it-1}^*) \neq 0$ and $E(s_{ist} y_{is}^* \Delta\varepsilon_{it}) = 0$. The former condition can be assumed, while the latter requires some work to show. A classical sufficient condition that ensures exogeneity is $E(\Delta\varepsilon_{it} | s_{ist}, y_{is}^*) = 0$. However, because $\Delta\varepsilon_{it}$, s_{ist}, y_{is}^* are related in a complicated way, it is not feasible to verify this condition in our context. A simpler sufficient condition derived in the Appendix is the following

$$E(d_{it} d_{it-1} d_{it-2} \Delta\varepsilon_{it} | d_{is}, y_{is}^*) = 0 \tag{18}$$

To see that this condition holds, substitute into $\Delta\varepsilon_{it}$ and write

$$E(d_{it}d_{it-1}d_{it-2}\Delta\varepsilon_{it}|d_{is},y^*_{is}) = E(d_{it}d_{it-1}d_{it-2}(\Delta\varepsilon^0_{it}+\vartheta_0\Delta u_{it})|d_{is},y^*_{is}) \tag{19}$$

$$= E(d_{it}d_{it-1}d_{it-2}\vartheta_0(u_{it}-u_{it-1})|d_{is},y^*_{is}) \tag{20}$$

because $\Delta\varepsilon^0_{it}$ is independent of d_{it}, d_{it-1}, d_{it-2}, d_{is}, and y^*_{is} and therefore it is independent of d_{it}, d_{it-1}, and d_{it-2}, conditionally on d_{is} and y^*_{is}. Now, conditioning additionally on η_i and d_{it-2},

$$E(d_{it}d_{it-1}d_{it-2}\Delta\varepsilon_{it}|d_{is},y^*_{is}) = \vartheta_0 E(d_{it-2}E(d_{it}d_{it-1}(u_{it}-u_{it-1})|\eta_i,d_{it-2},d_{is},y^*_{is})|d_{is},y^*_{is}) \tag{21}$$

notice that $d_{it}d_{it-1}(u_{it}-u_{it-1})$ is independent of d_{it-2}, d_{is}, and y^*_{is} conditionally on η_i. Therefore, $E(d_{it}d_{it-1}(u_{it}-u_{it-1})|\eta_i,d_{it-2},d_{is},y^*_{is}) = E(d_{it}d_{it-1}(u_{it}-u_{it-1})|\eta_i)$. It suffices then to show that $E(d_{it}d_{it-1}(u_{it}-u_{it-1})|\eta_i)=0$. Using conditional independence again, we obtain

$$E\left(d_{it}d_{it-1}(u_{it}-u_{it-1})|\eta_i\right) = E\left(d_{it}d_{it-1}u_{it}|\eta_i\right) - E\left(d_{it}d_{it-1}u_{it-1}|\eta_i\right) \tag{22}$$

$$= E\left(d_{it}u_{it}|\eta_i\right)E\left(d_{it-1}|\eta_i\right) - E\left(d_{it}|\eta_i\right)E\left(d_{it-1}u_{it-1}|\eta_i\right) = 0 \tag{23}$$

because $E\left(d_{it}u_{it}|\eta_i\right) = E\left(d_{it-1}u_{it-1}|\eta_i\right)$ and $E\left(d_{it}|\eta_i\right) = E\left(d_{it-1}|\eta_i\right)$ by the identical distributedness assumption. We have proven that

$$E(s_{ist}y^*_{is}\Delta\varepsilon_{it}(\rho)) = (\rho_0-\rho)E\left(s_{ist}y^*_{is}\Delta y^*_{it-1}\right) \tag{24}$$

Thus, we will have identification if and only if $E\left(s_{ist}y^*_{is}\Delta y^*_{it-1}\right)\neq 0$, that is, the same identification restriction as in the AB setting, except that here attention is restricted to observed data.

2.4.2 Bound on system estimator bias

Consider the infeasible level moment conditions $E((y^*_{it}-\rho_0 y^*_{it-1})\Delta y^*_{it-1})=0$. The feasible analogue for the moment on the left hand side is $E(d_{it}d_{it-1}d_{it-2}(y^*_{it}-\rho_0 y^*_{it-1})\Delta y^*_{it-1})$. However, we have verified in Monte Carlo experiments that it is not generally equal to zero in our model. Using the system moments anyway introduces a bias proportional to $E(d_{it}d_{it-1}d_{it-2}(\alpha_i+\varepsilon_{it})\Delta y^*_{it-1})$. Because we know that $E((\alpha_i+\varepsilon_{it})\Delta y^*_{it-1})=0$, we can write

$$E(d_{it}d_{it-1}d_{it-2}(\alpha_i+\varepsilon_{it})\Delta y^*_{it-1}) = E((1-d_{it}d_{it-1}d_{it-2})(\alpha_i+\varepsilon_{it})\Delta y^*_{it-1}) \tag{25}$$

Now, using the Cauchy-Schwarz inequality, we can write

$$|E(d_{it}d_{it-1}d_{it-2}(\alpha_i+\varepsilon_{it})\Delta y^*_{it-1})| \leq (1-P(d_{it}=d_{it-1}=d_{it-2}=1))\sqrt{\text{var}((\alpha_i+\varepsilon_{it})\Delta y^*_{it-1})} \quad (26)$$

Therefore, as selection disappears and every three consecutive values of y^*_{it} become observable, the bias goes to zero. Unfortunately, this is not a tight bound (i.e., it does not describe a worst-case scenario for the bias), so it is difficult to interpret the second factor in the bound. The Cauchy-Schwarz inequality $E|XY| \leq \sqrt{EX^2}\sqrt{EY^2}$ holds with equality if and only if X and Y are proportional. Because $1-d_{it}d_{it-1}d_{it-2}$ is discrete, it cannot possibly be proportional to the continuously distributed $(\alpha_i+\varepsilon_{it})\Delta y^*_{it-1}$.

2.5 Consistency in the dynamic model with covariates

2.5.1 An exogenous covariate

We extend the previous AR(1) model to a model with a single exogenous covariate (although the result can be straightforwardly generalised to many covariates),

$$y^*_{it} = \alpha_i + \rho_0 y^*_{it-1} + \beta_0' x^*_{it} + \varepsilon_{it} \quad (27)$$
$$d_{it} = 1(\eta_i + \gamma_0 z_{it} + u_{it} > 0) \quad (28)$$
$$\alpha_i = \alpha^0_i + \theta_0 \eta_i \quad (29)$$
$$\varepsilon_{it} = \varepsilon^0_{it} + \vartheta_0 u_{it} \quad (30)$$

The exogenous random variables x^*_{it}, z_{it}, α^0_i, ε^0_{it}, η_i, and u_{it} are assumed to be i.i.d. and independent of each other with finite second moments.[5] We assume that $|\rho|<1$ and y^*_{it} is the stationary causal solution to the AR(1) model, $y^*_{it} = \frac{\alpha_i}{1-\rho_0} + \sum_{j=0}^{\infty}\rho_0^j(\beta_0' x^*_{it} + \varepsilon_{it-j})$. We also assume that $E(\varepsilon^0_{it}) = E(u_{it}) = 0$. The observed data is the set of y^*_{it} and x^*_{it} for which $d_{it}=1$.

Now, define $\Delta\varepsilon_{it}(\rho,\beta) = \Delta y^*_{it} - \rho\Delta y^*_{it-1} - \beta'\Delta x^*_{it}$ and write

$$E(s_{ist}y^*_{is}\Delta\varepsilon_{it}(\rho,\beta)) = (\rho_0-\rho)E(s_{ist}y^*_{is}\Delta y^*_{it-1}) + (\beta_0-\beta)'E(s_{ist}y^*_{is}\Delta x^*_{it}) + E(s_{ist}y^*_{is}\Delta\varepsilon_{it}) \quad (31)$$
$$E(s_{ivt}x^*_{iv}\Delta\varepsilon_{it}(\rho,\beta)) = (\rho_0-\rho)E(s_{ivt}x^*_{iv}\Delta y^*_{it-1}) + (\beta_0-\beta)'E(s_{ivt}x^*_{iv}\Delta x^*_{it}) + E(s_{ivt}x^*_{is}\Delta\varepsilon_{it}) \quad (32)$$

It is clear that identification requires that for some t and some v, the matrix

$$\begin{bmatrix} E(s_{ist}y^*_{is}\Delta y^*_{it-1}) & E(s_{ist}y^*_{is}\Delta x^*_{it}) \\ E(s_{ivt}x^*_{iv}\Delta y^*_{it-1}) & E(s_{ivt}x^*_{iv}\Delta x^*_{it}) \end{bmatrix}$$

[5] We use x^*_{it} to note that, even in the case of assuming exogeneity, the covariate could also be partially unobserved.

is non-singular.

We have already shown that $E(s_{ist}y^*_{is}\Delta\varepsilon_{it}) = 0$. It remains to show that $E(s_{ivt}x^*_{iv}\Delta\varepsilon_{it}) = 0$. Now,

$$\begin{aligned}
E(s_{ivt}x^*_{iv}\Delta\varepsilon_{it}) &= E(d_{it}d_{it-1}d_{it-2}d_{iv}x^*_{iv}(\Delta\varepsilon^0_{it} + \vartheta_0\Delta u_{it})) && (33)\\
&= E(d_{it}d_{it-1}d_{it-2}d_{iv}x^*_{iv}\vartheta_0\Delta u_{it}) && (34)\\
&= E(d_{it-2}d_{iv}x^*_{iv}\vartheta_0 E(d_{it}d_{it-1}\Delta u_{it}|\eta_i, d_{it-2}, d_{iv}, x^*_{iv})) && (35)\\
&= E(d_{it-2}d_{is}x^*_{iv}\vartheta_0 E(d_{it}d_{it-1}\Delta u_{it}|\eta_i)) && (36)\\
&= 0. && (37)
\end{aligned}$$

The first equality follows from the independence of ε^0 from all other variables. The second equality is obtained by conditioning on predetermined variables. The third equality follows from the conditional independence of $d_{it}d_{it-1}\Delta u_{it}$ from $(d_{it-2}, d_{is}, x_{is})$ conditional on η_i. The final equality has already been established above.

2.5.2 A predetermined covariate

Now, suppose that x^* is predetermined so that x^*_{it} is independent of $\varepsilon^0_{it+1}, \varepsilon^0_{it+2}, \ldots$, $u_{it+1}, u_{it+2}, \ldots$, and $z_{it+1}, z_{it+2}, \ldots$ but not necessarily independent of contemporaneous or past values of these variables. Then, exogeneity may still be satisfied if $v \leq t-2$ in the above calculations. If we can further assume that x_{iv} is independent of ε_{iv}, u_{iv}, and z_{iv}, then exogeneity will be satisfied with $v = t-1$ as well.

2.5.3 An endogenous covariate

Finally, suppose x^* is endogenous and we have at our disposal a vector of instruments ξ. Then, we may use the following moment conditions

$$\begin{aligned}
E(s_{ist}y^*_{is}\Delta\varepsilon_{it}(\rho,\beta)) &= (\rho_0-\rho)E(s_{ist}y^*_{is}\Delta y^*_{it-1}) + (\beta_0-\beta)'E(s_{ist}y^*_{is}\Delta x^*_{it}) + E(s_{ist}y^*_{is}\Delta\varepsilon_{it}) && (38)\\
E(s_{it}\xi_i\Delta\varepsilon_{it}(\rho,\beta)) &= (\rho_0-\rho)E(s_{it}\xi_i\Delta y^*_{it-1}) + (\beta_0-\beta)'E(s_{it}\xi_i\Delta x^*_{it}) + E(s_{it}\xi_i\Delta\varepsilon_{it}), && (39)
\end{aligned}$$

where $s_{it} = d_{it}d_{it-1}d_{it-2}$. Thus, we need

$$\begin{bmatrix} E(s_{ist}y^*_{is}\Delta y^*_{it-1}) & E(s_{ist}y^*_{is}\Delta x^*_{it}) \\ E(s_{ivt}x^*_{iv}\Delta y^*_{it-1}) & E(s_{ivt}x^*_{iv}\Delta x^*_{it}) \end{bmatrix}$$

to be non-singular, and we need $E(s_{ist}y^*_{is}\Delta\varepsilon_{it}) = 0$ and $E(s_{it}\xi_i\Delta\varepsilon_{it}) = 0$.

3 Monte Carlo experiment of the pure AR(1) model

For the Monte Carlo experiment, we consider the following data-generating processes. First, we assume two different options for the selection equation:

$$d_{it}^{*} = a - z_{it} - \eta_i - u_{it} \tag{40}$$

$$d_{it}^{*} = a - 0.5d_{it-1} + z_{it} - \eta_i - u_{it} \tag{41}$$

$$d_{it} = 1[d_{it}^{*} > 0] \tag{42}$$

where a is set so $p(d_{it}^{*} > 0) = 0.85$ and $z_{it} \sim N(0, \sigma_z)$ with $\sigma_z = 1$. Second, the outcome of interest is generated as follows:

$$y_{it}^{*} = (2 + \alpha_i + \varepsilon_{it})/(1 - \rho) \text{ if } t = 1 \tag{43}$$

$$y_{it}^{*} = 2 + \rho y_{it-1}^{*} + \alpha_i + \varepsilon_{it} \text{ if } t = 2, ..., T \tag{44}$$

$$y_{it} = y_{it}^{*} \text{ if } d_{it} = 1 \tag{45}$$

We let ρ vary between 0.25, 0.50 and 0.75. We generate all variables for $T = 17$ to $T = 20$ and discard the first 13 observations to minimise any problem with initial conditions.[6] Finally, we assume the following structure for the errors:

$$\eta_i \sim N(0, \sigma_\eta) \text{ with } \sigma_\eta = 1 \tag{46}$$

$$u_{it} \sim N(0, \sigma_u) \text{ with } \sigma_u = 1 \tag{47}$$

$$\alpha_i = \alpha_i^0 + 0.5\eta_i, \alpha_i^0 \sim N(0, \sigma_{\alpha^0}) \text{ with } \sigma_{\alpha^0} = 1 \tag{48}$$

$$\varepsilon_{it} = \varepsilon_{it}^0 + 0.5u_{it}, \varepsilon_{it}^0 \sim N(0, \sigma_{\varepsilon^0}) \text{ with } \sigma_{\varepsilon^0} = 1 \tag{49}$$

These assumptions imply that $corr(\varepsilon_{it}, u_{it}) = corr(\alpha_i, \eta_i) = 0.5/\sqrt{1 + 0.5^2} = 0.447$.

3.1 Description of the experiments

For each experiment, we set the initial sample size to $N = 500$ or $N = 5000$, and for each i, we draw up to 20 time series observations, from which the initial 13 are discarded. Once selection is applied, the unbalanced panels are formed. At least three consecutive observations of the same regime are needed to form an observation of the selected panel. This implies that a large fraction of the observations do not contribute to the identification of the parameters, even with a small

[6] However, the results remain unchanged if we do generate these extra 13 observations and, thus, start the observed sample with an initial condition for each individual in the sample.

degree of sample selection. For example, a 15 per cent of initial selection implies that around 1/3 of the observations are lost. For each combination of the parameters we perform 500 replications.

In each case, we evaluate the performance of two well-known GMM-IV estimators: AB and system. The structure of the model makes selection of the instruments a crucial step of this simulation study. We select the instruments as follows: we use lags from $t-2$ backwards for first-differenced equations, although we also evaluate the performance of the estimates with a restricted set of instruments. We use the lagged first difference of the outcome as an additional instrument for the equation in levels. Although we are aware of the instrument proliferation issue analysed by Roodman (2009), it does not constitute a problem here given the reduced number of periods (a maximum of 7) remaining for estimation.[7]

3.2 Simulation results for the pure autoregressive model

3.2.1 The basic results

Table **1** presents results for the AR(1) model for three values of the autoregressive parameter: 0.25, 0.50 and 0.75.[8] We simulate two alternative selection models (static, A, and dynamic, B), as presented in equations (40) and (41). For each combination of selection model and autoregressive parameter, we report results for both the AB and the system estimators constructed under competing assumptions about the selection process: (a) non-endogenous selection; (b) endogenous selection without correction. The initial degree of sample selection is 15 per cent, while the fraction of the sample lost is much larger (around 1/3 of the observations).

Let us start reviewing the results without endogenous selection, reported in columns (1) and (2). When the initial sample (before selecting the observations) is small ($N = 500$) the bias of the AB grows with the autoregressive parameter (for both selection models, A and B) and becomes sizable when $\rho = 0.75$.[9] As we increase the sample size ($N = 5000$), the average bias of the AB estimator is reduced substantially and only remains noticeable for $\rho = 0.75$. Alternatively, the system estimator, which is also consistent in this case, shows a very small bias for $N = 500$ (never exceeding one per cent), even smaller when $N = 5000$. Figure **1** confirms these results with a sample size varying from $N = 200$ to $N = 5000$ in absence of any sort of selection (estimators labelled AB all and system all).

When endogenous sample selection is considered (see columns (3) and (4) for AB and system estimators results) but we do not include any correction for endogenous sample selection in the model, we do not detect any significant change in the bias results for the AB estimator for both selection models, even when the initial sample is small . In fact, when the initial sample is small,

[7] We used Roodman's proposal to collapse the number of instruments and we get very similar results in the empirical application.

[8] Results for other values of the autoregressive parameter are available upon request. For example, for values below 0.25 (for example, 0.10), the results remain unchanged. For values closer to one (for example, 0.90), the bias is larger but not worse than the one found in, for example, the balanced sample.

[9] See Blundell and Bond (1998) and Hayawaka (2007) for analyses of the small sample bias of the AB and system GMM estimators in linear models.

the difference between the cases with and without selection is practically undetectable (see Table A1). Alternatively, the results confirm the consistency of the AB and the small bias of the system estimator for $N = 5000$. In contrast, the system estimator always shows a very small bias (between 1 per cent for $\rho = 0.25$ and 2.25 per cent for $\rho = 0.75$). Note that the bias becomes more evident as the sample size grows (see Figure **1**). As a sort of compensation, the standard errors always tend to be substantially smaller.

Some additional conclusions can be drawn when varying sample size (Figure **1**). When $N = 200$, the AB estimator shows sizable bias, which decreases as N increases. The system estimator has a very small bias, however. For a given ρ, it remains stable (between 1 and 2.5 per cent) as N increases. We can find a threshold for N for each combination of parameters. Below this threshold, the average bias of the system estimator is smaller, and it is larger above it. Therefore, we may conclude that for moderate and small samples (say, below the range 1000-1500), the system estimator is highly recommended because of the likely smaller small sample bias as well as smaller variance.

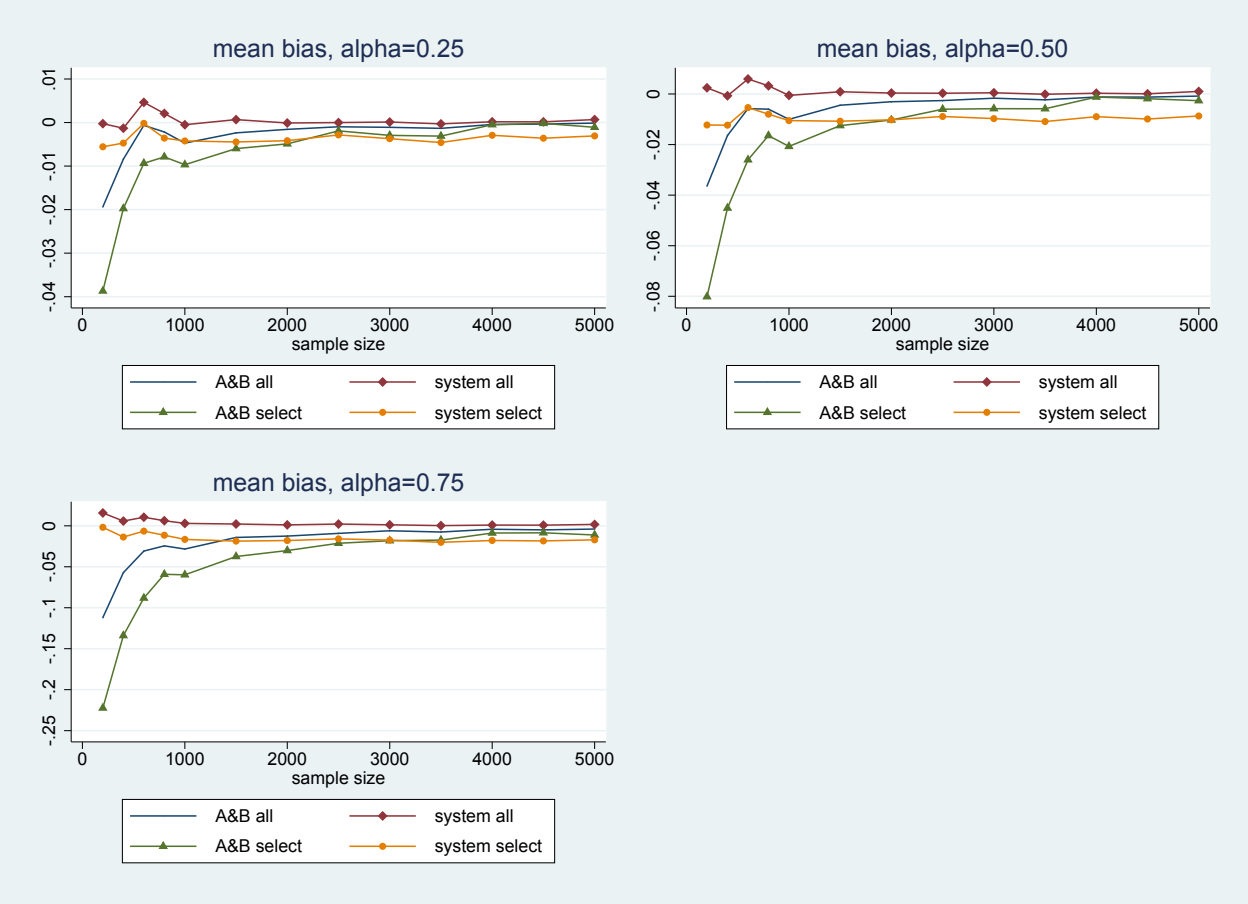
Finally, as shown in Figure **2**, when the individual heterogeneity components, α_i and η_i, are not correlated, the bias of the system estimator practically disappears (in comparison with the previous case) due to the fact that the main source of bias is the correlation between the heterogeneous components of both the outcome and selection equations (see Table A.1 for an illustration).[10] This means that in those cases in which endogenous selection is not due to individual heterogeneity, all three estimators considered are, in essence, valid options for recovering the key parameters of the outcome equation.

[10]Table A.1 in the Appendix presents an analysis of the conditional expectation of the key moment conditions of the model for different values of N, ρ and correlations between the error components and the autoregressive parameter.

Table 1: **Average bias in the AR(1) model ($T = 7$, 500 replications)**

			No endogenous selection		Endogenous selection	
Select. Model	ρ		(1) AB	(2) SYS	(3) AB	(4) SYS
			$N = 500$			
A	.25	bias	-.01002	.00302	-.01054	-.00151
		s.e.	.06004	.04378	.05447	.04209
A	.50	bias	-.03265	.00224	-.02964	-.00700
		s.e.	.08833	.04990	.07807	.05027
A	.75	bias	-.19339	.00757	-.10176	-.00923
		s.e.	.19402	.05879	.13350	.06437
			$N = 5000$			
A	.25	bias	-.00209	.00017	-.00059	-.00381
		s.e.	.01661	.01272	.04358	.02387
A	.50	bias	-.00461	.00019	-.00264	-.00868
		s.e.	.02479	.01491	.02285	.01493
A	.75	bias	-.02395	.00112	-.01118	-.01709
		s.e.	.05792	.01791	.03863	.01894
			$N = 500$			
B	.25	bias	-.01076	.00216	-.01075	-.00024
		s.e.	.05904	.04293	.05471	.04219
B	.50	bias	-.03324	.00135	-.03022	-.00556
		s.e.	.08802	.04925	.07843	.05039
B	.75	bias	-.18966	.00709	-.10478	-.00823
		s.e.	.19100	.05824	.13880	.06358
			$N = 5000$			
B	.25	bias	-.00208	-7.75e-07	-.00103	-.00134
		s.e.	.01680	.01267	.01700	.01279
B	.50	bias	-.00459	.00019	-.00263	-.00649
		s.e.	.02540	.01481	.02380	.01482
B	.75	bias	-.02375	.00105	-.01129	-.01472
		s.e.	.05964	.01780	.04192	.01871

Figure 1: Average bias of the AB and system estimators in the full sample ($(NxT$ observations) and the endogenously selected sample



Notes.

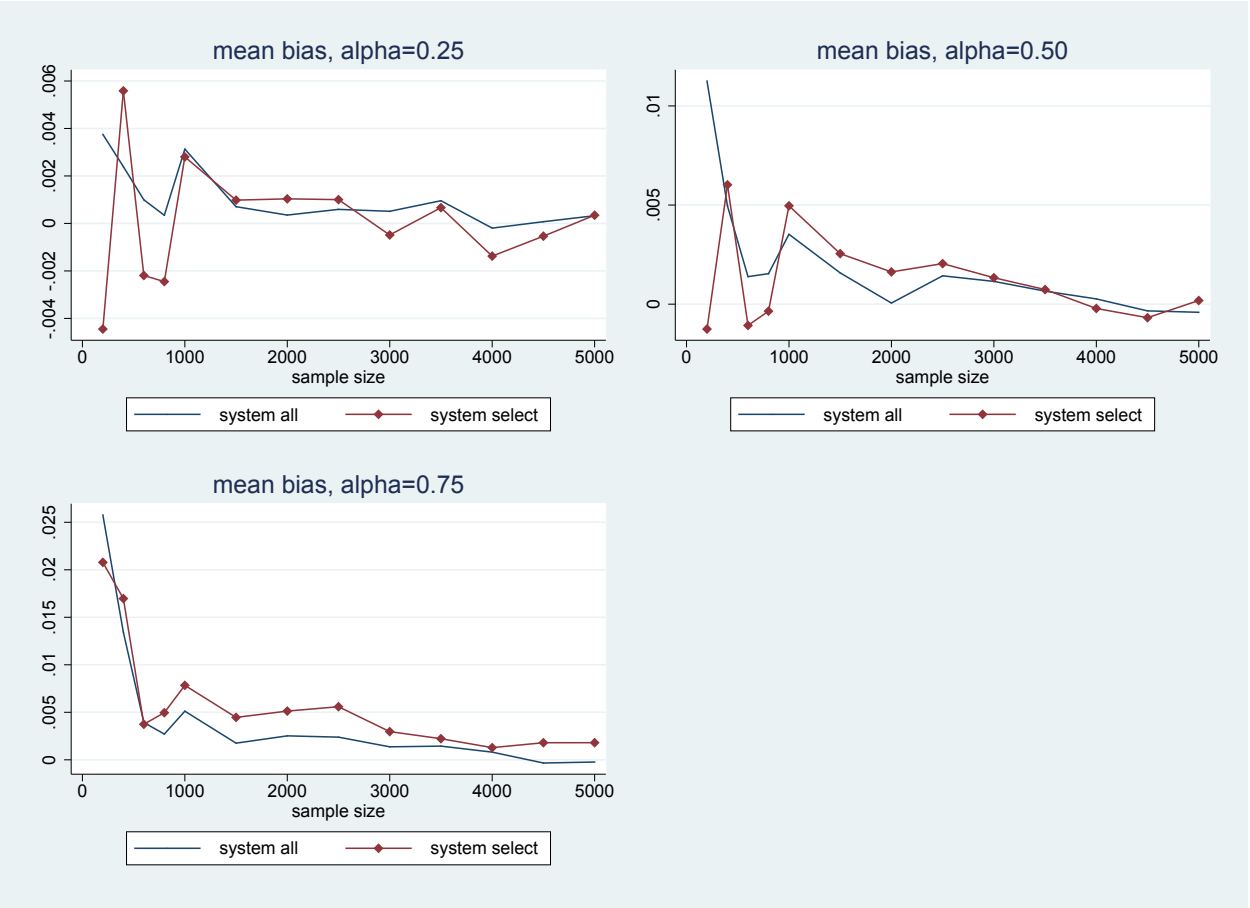
AB all: AB GMM-IV estimates using the full (NxT) sample (no selection process).

system all: System GMM-IV estimates using the full (NxT) sample (no selection process).

AB select: Uncorrected for selection AB GMM-IV estimates on the selected sample under endogenous sample selection.

system select: Uncorrected system GMM-IV estimates on the selected sample under endogenous sample selection.

Figure 2: **Average bias of system estimator in the full sample (NxT observations) and the endogenously selected sample when α_i and η_i are not correlated**



Notes.

system all: System GMM estimates with the full sample (no-selection).

system select: Uncorrected system GMM estimates with the selected sample under endogenous selection due to correlation of the time-varying errors.

3.2.2 Sensitivity analysis

In this section, we comment on various departures from the basic assumptions. We consider the following representative cases: (a) varying the longitudinal dimension of the panel; (b) increasing the percentage of selection (from 0.15 to 0.25); (c) increasing the ratio of the variances to $\frac{\sigma_\alpha^2}{\rho_\varepsilon^2} = 2$; (d) reducing the correlation between the errors (the correlation parameter is reduced from 0.5 to 0.25); (e) and, finally, non-stationary time varying errors and correlation of the time-varying error components. In particular, we allow the variance of the time-varying errors in (1) and (2) to vary over time[11] and we also allow the correlation coefficient between the time-varying errors in (1) and (2) to vary over time.[12] We present the simulation results for $N = 500$ in Table **2** and for $N = 5000$

[11] We multiply either ε_{it} or u_{it} by a time-varying Bernoulli process taking either 1 or 2.

[12] We multiply ϑ by either 0.5, 1 or 2.

in Table **3**, for three values of the autoregressive parameters: 0.25, 0.50 and 0.75.

Our first experiment reduces the maximum longitudinal dimension of the observed panel from $T = 7$ to $T = 4$. Apart from the expected increase in the estimated variance and regardless of the sample size considered, the effect on the average bias of the AR(1) coefficient implied by this change is very small for both estimators, all values of the autoregressive parameter considered and both sample sizes.

Increasing the degree of sample selection from 0.15 to 0.25 increases average bias of the autoregressive coefficient very mildly. In addition, it increases its variance due to the significant reduction in the number of observations (the average number of observations is reduced around 30 per cent).

The increase of the ratio of the variance of the individual heterogeneous component to the variance of the time-variant component of the outcome equation does not have an important effect on the average bias of the estimated parameters, either in the AB or the system GMM estimator. We can observe in Tables **2** and **3** that the effect is smaller when the sample size is larger.

We also consider a reduction in the correlation parameter of the errors. In particular, we assume the following structure for the errors: $\varepsilon_{it} = \varepsilon_{it}^0 + 0.25u_{it}$ and $\alpha_i = \alpha_i^0 + 0.25\eta_i$, which implies a correlation coefficient of $0.2425(= 0.25/\sqrt{1 + 0.25^2})$ in either case. As can be easily detected by comparing the results reported in Tables **1**, **2** and **3**, this change reduces the average bias of the estimators for both sample sizes and all autoregressive parameters. Finally, when we introduce in the model non-stationary time-varying error components and we allow the correlation between the errors in the outcome and the selection equation to vary over time, the average bias of the estimated parameters is significantly reduced, specially important when the initial sample size is small.

In sum, these sensitivity exercises confirms the main lessons we can draw from the analysis: (a) the AB (or the AH) estimator is moderately biased when N is small or moderate, and unbiased when N is large. The system GMM estimator is always moderately biased. All these results imply that the system estimator is especially recommended when the sample size is small or even moderate (below one or one and a half thousand individuals) and less "important" when the sample size is larger.

Table 2: **Average bias in the AR(1) model. Sensibility analysis for small N**

Model		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
		AB	SYS	AB	SYS	AB	SYS
Experiment I: Very short T ($T = 4$)							
A	bias	-.00351	-.00096	-.01775	-.01349	-.08907	-.03272
	s.e.	.13394	.07796	.20964	.09687	.39245	.13551
B	bias	-.00094	.00299	-.01147	-.00949	-.07759	-.02922
	s.e.	.13261	.07737	.21371	.09551	.44549	.13314
Experiment II: More sample selection (25%)							
A	bias	-.01878	-.00135	-.04518	-.00622	-.13019	-.00601
	s.e.	.06665	.05032	.09556	.06172	.15362	.07985
B	bias	-.01941	.00122	-.04817	-.00377	-.14386	-.00476
	s.e.	.06811	.05026	.09801	.06047	.16296	.07769
Experiment III: Increasing the ratio of variances: $\sigma_\eta/\sigma_\epsilon = 2$							
A	bias	-.01095	.00071	-.03149	-.00263	-.11253	.00535
	s.e.	.05608	.04446	.08088	.05430	.14306	.07006
B	bias	-.01121	.00162	-.03211	-.00151	-.11640	.00565
	s.e.	.05596	.04433	.0812	.05433	.14816	.06920
Experiment IV: Reducing the correlation of the errors: $\rho = 0.25$							
A	bias	-.01018	.00109	-.03182	-.00069	-.14046	.00340
	s.e.	.05720	.04293	.08392	.04999	.15923	.06081
B	bias	-.01077	.00077	-.03197	-.00050	-.14149	.00416
	s.e.	.05673	.04243	.08376	.04961	.15990	.06031
Experiment V: Non-stationary time-varying error components							
A	bias	-.01263	-.00333	-.02599	-.00784	-.07935	-.00925
	s.e.	.05537	.03968	.07605	.04742	.12415	.05987
B	bias	-.01246	-.00179	-.02854	-.00593	-.07977	-.00781
	s.e.	.05354	.03962	.07549	.04438	.12231	.05692

Notes.

1. $T = 7$, except in experiment I.
2. $N = 500$.
3. Number of replications: 500.

Table 3: **Average bias in the AR(1) model. Sensibility analysis for large N**

Model		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
		AB	SYS	AB	SYS	AB	SYS
Experiment I: Very short T ($T = 4$)							
A	bias	-.00059	-.00381	-.00118	-.01070	-.00258	-.02123
	s.e.	.04358	.02387	.06645	.02801	.11215	.03444
B	bias	-.00006	-.00191	-.00106	-.00858	-.00413	-.01891
	s.e.	.04311	.02393	.06637	.02797	.11311	.03412
Experiment II: More sample selection (25%)							
A	bias	-.00226	-.00402	-.00462	-.01033	-.01433	-.01962
	s.e.	.02124	.01549	.02968	.01774	.04639	.02241
B	bias	-.00218	-.00151	-.00463	-.00724	-.01566	-.01658
	s.e.	.02158	.01572	.03027	.01803	.04780	.02257
Experiment III: Increasing the ratio of variances: $\sigma_\eta/\sigma_\epsilon = 2$							
A	bias	-.00123	-.00248	-.00295	-.00721	-.01326	-.01454
	s.e.	.01720	.01351	.02400	.01608	.04137	.02094
B	bias	-.00108	-.00099	-.00288	-.00539	-.01323	-.01253
	s.e.	.01736	.01348	.02493	.01606	.04494	.02076
Experiment IV: Reducing the correlation of the errors: $\rho = 0.25$							
A	bias	-.00165	-.00078	-.00373	-.00217	-.01594	-.00370
	s.e.	.01653	.01275	.02423	.01506	.04627	.01838
B	bias	-.00164	-.00030	-.00368	-.00159	-.01615	-.00299
	s.e.	.01679	.01264	.02489	.01499	.04916	.01812
Experiment V: Non-stationary time-varying error components							
A	bias	-.00219	-.00246	-.00142	-.00586	-.01400	-.01263
	s.e.	.01923	.01288	.03185	.01615	.05631	.02099
B	bias	-.00197	-.00089	-.00306	-.00484	-.01688	-.01152
	s.e.	.01975	.012491	.02874	.01535	.05961	.02037

Notes.

1. $T = 7$, except in experiment I.
2. $N = 5000$.
3. Number of replications: 500.

4 An evaluation of simple alternatives for bias correction of the system estimator

Bias correction induced by endogenous sample selection implies adding univariate corrections if the sample is only conditional on one observation (random effects strategy in the static model) as shown by Wooldrige (1995), and bivariate corrections conditional on two periods (first differences fixed effects strategy in the static model) as shown by Rochina-Barrachina (1999). However, we have shown that IV and GMM-IV estimators of the model in first differences do not require any

correction to be consistent. Furthermore, in the case of the system GMM-IV estimator, we have shown that only the equations in levels need to be corrected. In addition, we have shown that the bias of the additional moment conditions implied by the use of equations in levels is caused by the correlation between the unobserved time-invariant heterogeneity components in the outcome and selection equations (see Table A1 for an illustration). In what follows, we describe simple procedures to account for each of these problems. We focus on a static selection equation but we will also introduce some comments about the dynamic case.

4.1 An univariate correction for the system estimator

As previously mentioned, the correction for selection is only needed in the equations in levels. When we also introduce corrections in the equations in first differences, it should not affect identification of the rest of the parameters. In this section, following Wooldridge (1995) combined with a fixed effect strategy, we consider univariate corrections. Irrespective of whether we correct both the level and the first-differenced equations or the the level ones anole, the procedure, following Wooldridge (1995), can be described as follows:

- Step 1. Estimate year-by-year probit models (under normality) for either the static or dynamic selection models following the Mundlak/Chamberlain/Wooldridge approach and compute univariate correction terms (Heckman's lambda).

- Step 2. Estimate the corrected equations by GMM-IV, adding, to either the equation in levels or all the equations, the selection term as an additional regressor.

  In the case of correcting both levels and the first-differenced equations, standard software can be used (see, for instance, Roodman, 2006). Alternatively, when only the equations in levels are corrected, the system estimator can be obtained using, for instance, the Stata ***gmm*** routine. (Robust) Corrected standard errors need to be computed anyway. This can be done by means of the delta method or bootstrapping.[13]

Finally, a standard t-test of significance of the correction term (or a Wald test, in case of letting the effect of the correction to vary over time) stands for an approximate test of endogenous selection (Wooldridge, 1995).

4.1.1 Construction of the correction

For a typical static selection model, as described in equation (2), and assuming normality of $\eta_i + u_{it}$, we estimate a probit for each period and then compute the well-known selection term $\hat{\lambda}_{it}(z_{it}\hat{\gamma})$. When we allow correlation between z_{it} and η_i, we can rely on Mundlak (1978) and assume, for instance, $\eta_i = \tilde{z}_i\varphi$, where $\tilde{z}_i$ is the vector of individual means of z_{it}, and we, again, can estimate a

[13] See the Appendix for a proposal to correct the variance of the (corrected) GMM estimators following Terza (2016).

probit for each period and compute $\tilde{\lambda}_{it}(z_{it}\tilde{\gamma} + \tilde{z}_i\tilde{\varphi})$, which is then introduced in a second step as before.

In the case of a dynamic selection equation, the lagged observed regressor is correlated with the random effect by construction. Is this is the case we need to rely either on Mundlak's proposal or on a less restrictive one such as that of Chamberlain (1984). In the latter case, we can assume $\eta_i = \pi_1 z_{i1} + \pi_2 z_{i2} + ... + \pi_T z_{iT}$ and recover the corresponding selection terms.[14]

4.2 A simple procedure for bias reduction of the system estimator in the presence of endogenous selection

As stated before and shown in Figure **2**, a large fraction of the inconsistency of the system estimator stems from the correlation between the unobserved heterogeneous components in equations (1) and (2). Because many practitioners are potentially interested in estimating these models using the system estimator (especially when the sample size is small), we describe a simple procedure to obtain it, and we also suggest a test. The procedure can be described as follows:

- Step 1: Provided the selection equation has an exclusion restriction, obtain a consistent estimate of the fixed effects (say, $\hat{\eta}_i$) in the selection equation using a linear probability model.

- Step 2: Add the estimate of η_i to the equation in levels to control the correlation between the time-invariant errors.

$$y_{it} = \rho y_{it-1} + \alpha_i^* + \theta\hat{\eta}_i + \epsilon_{it} \quad for\, t_i \quad s.t.\, d_{it}, d_{it-1}, d_{it-2} = 1$$

  now α_i^* is purged of any correlation with the time invariant component in the selection equation.

- Step 3: Obtain the system estimator combining the uncorrected equations in first differences and the corrected equations in levels (with the corrections mentioned above).

A simple t-test of the null $\theta = 0$ stands for a test of endogenous selection. As in the previous case, corrected standard errors can be computed using the delta method or bootstrapping. If we cannot reject the null hypothesis, the individual heterogeneity components are uncorrelated, so the only potential source of endogenous selection is the correlation of the time-variant errors. Therefore, the only remaining problem for the consistency of the system GMM estimator is the

[14] Strictly speaking, to recover the structural parameters of the selection equation, we should estimate a probit for each year based on a reduced form, where d_{it}^* is modeled as a function of all exogenous variables (the $z's$) and we predict the index $\hat{d}_{it}^*$. Then, in a second stage, we estimate the structural parameters by within-groups, MD or GMM and compute the correction terms based on these two-stage estimates (see Bover and Arellano, 1997, or Labeaga, 1999). However, to keep the exercise as simple as possible, we compute the selection terms using reduced-form estimates for each period.

potential correlation between the time varying errors of both equations. However, we show in Table A.1 that this correlation is not inducing, in general, much bias.

4.3 Evaluation of the proposals

Figure **3** compares the average bias results of the uncorrected system estimator and two corrected ones: the year-by-year correction and the individual heterogeneity correction described above in step 2. As in previous figures, we let the initial sample size vary from 200 to 5000, $T = 7$ and let ρ take the values 0.25, 0.50 and 0.75. Furthermore, the probability of selection is 0.15 and the correlation between the error components 0.447.

The uncorrected system GMM estimator is moderately biased in all cases. The bias stabilizes as N grows to 1 per cent for $\rho = 0.25$ and to 2.5 per cent for $\rho = 0.75$. The average bias of the year-by-year corrected system estimator barely improves the one observed in the uncorrected case. We obtain the same result when considering either bivariate (including current and lagged lambda) or trivariate corrections (including current, lagged and twice-lagged lambda).

In stark contrast with the poor performance of the correction above, the system estimator obtained including the heterogeneous component adjusted in a linear probability model improves substantially from the uncorrected estimator. When $\rho = 0.25$, the average bias practically disappears regardless of the sample size. For higher values of ρ, it gets reduced substantially.

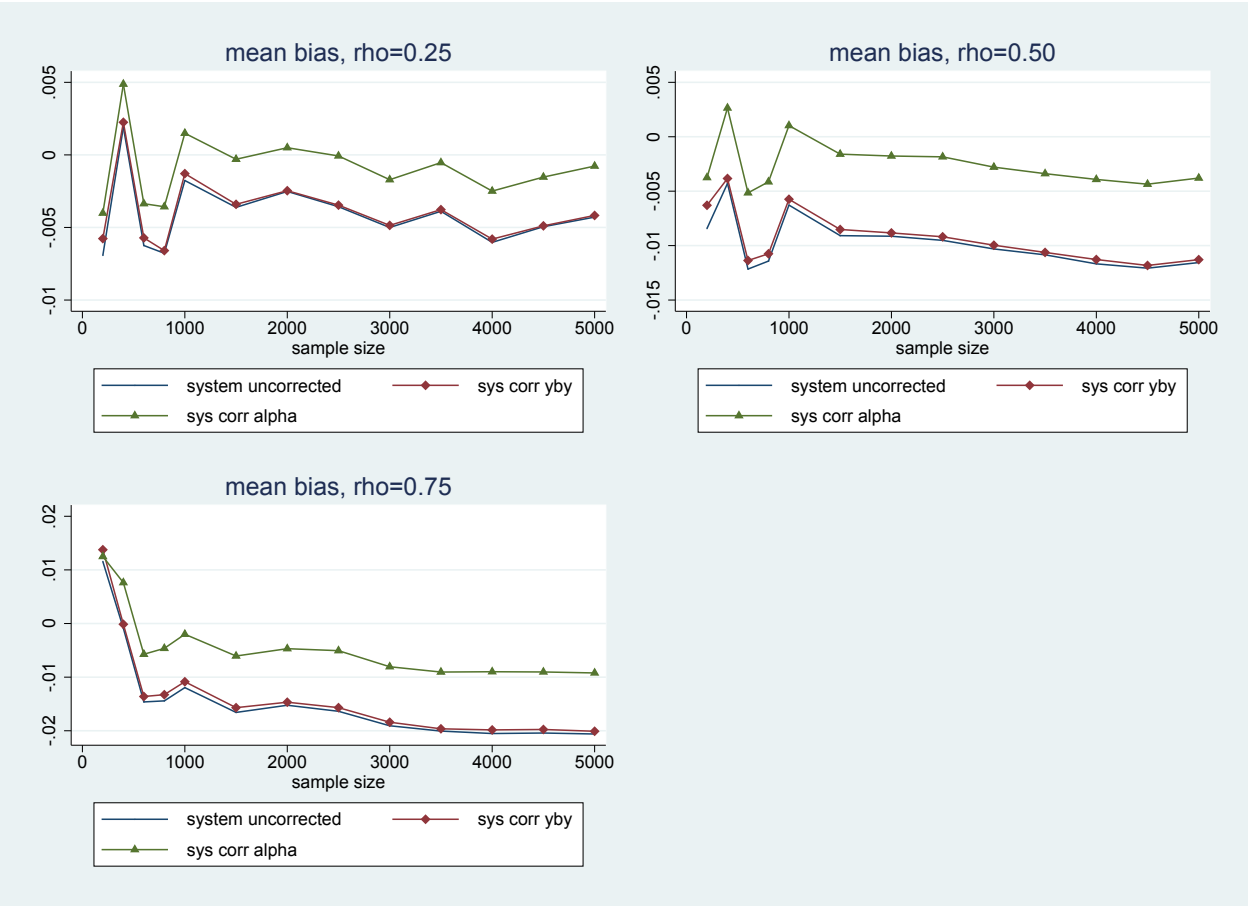
4.4 Evaluation of the tests

In Table **4**, we evaluate the size and power of the aforementioned tests. We keep the same assumptions of the previous subsection. We compute the empirical rejection frequency when the null is either true or false, choosing $\alpha = 0.05$ as the significance level. The test of the null $\theta = 0$ works well in terms of power, especially for large N, but not so well in terms of size, especially when the initial sample size is $N = 5000$, where the size exceeds the nominal size. Surprisingly, the test of the hypothesis $\gamma = 0$ works reasonably well in terms of both size and power, either with $N = 500$ or $N = 5000$.

Table 4: Empirical rejection frequency for the correction test, 500 reps.

		Correlated individual heterogeneity H_0: $\theta = 0$		Endogeneous selection H_0: $\gamma = 0$	
N	ρ	False power	True size	True power	False size
500	.25	.97	.076	.812	.05
500	.50	.962	.084	.81	.05
500	.75	.88	.076	.83	.05
5000	.25	1	.24	1	.06
5000	.50	1	.246	1	.06
5000	.75	1	.258	1	.054

Figure 3: Average bias of the system estimator. Simple alternatives for correction



Notes.

Probability of correction: 0.15.

Correlation between error components: 0.447.

System uncorrrected: uncorrected system GMM-IV.

Sys corr yby: system GMM-IV corrected using a year-by-year correction.

System corr alpha: system GMM-IV corrected accounting for the correlation between the time-invariant heterogeneity components.

5 Empirical Application

This section presents an application of the proposed methods. We employ the same data used in SW, which were also used by Lai and Tsai (2016).[15] The data consists of a panel taken from the Panel Study of Income Dynamics (PSID) covering the period 1980-1992, and we use the same selection rules (see Section 6 in Semykina and Wooldridge, 2013). The results for the pure autoregressive model are presented in Table **5**. Then, we extend the model in Table **6** to include age, age squared and level of education (number of years). The first column in Table **5** presents first-differenced IV estimates. Alternatively, column (1) in Table **6** reports the SW estimator. Columns (2) and (3) in both tables report AB and system results obtained in the selected sample, but when we do not correct the earnings equation. Column (4) adds a correction for the correlation between the unobserved heterogeneous components. Finally, in columns (5) and (6) of both tables, we present the system GMM results, adding time-varying correction terms estimated from year-by-year univariate probits. In column (5), we only correct the equation in levels, and in column (6), we correct both the levels and the first-differenced equations.[16]

Table 5: **AR(1) log hourly earnings equation**

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	2SLS-IV	*No correction*	*No correction*	*Het. components correction of lev eq. only*	*yby correction of lev eq. only*	*yby correction of all equations*
		AB	*system*	*system*	*system*	*system*
Lag log hourly earnings	0.1522**	0.1029**	0.1798***	0.1791***	0.2354***	0.2157***
	(0.0489)	(0.0377)	(0.0434)	(0.0436)	(0.0444)	(0.04397)
$\hat{\eta}_i$				0.0438		
				(0.0305)		
Observations	5033	5033	5033	5033	5033	5033
Joint significance selection terms					105.13 (11)	53.59 (11)
					(0.000)	(0.000)

Notes.

1. $N = 550$.
2. Annual dummies are included in all specifications.
3. * significant at 1%; ** significant at 5%; *** significant at 10%.

[15] We compare our results with those presented by SW, but, unfortunately, we cannot compare with Lai and Tsay (2016) because they estimated a static sample selection model.

[16] All the AB and system GMM estimates, except those reported in column (5), were obtained using the stata xtabond2 package (see Rodman, 2006). The estimates reported in column (5) have been obtained using a modified version of xtabond2 that only includes the correction in the level equations. Note, however, that these estimates can be also obtained using the Stata *gmm* routine.

4. The Standard Errors have been corrected following Windmeijer (2005). In columns (4) to (6), we also report corrected standard errors following Terza (2016). See the Appendix for details.

5. The test of significance of the selection terms is a Wald test. Degrees of freedom and level of significance are in parentheses.

In Table **6**, we consider the demographic variables to be strictly exogenous and we instrument the lagged log of the dependent variable using all available instruments for both the equations in levels and first-differences. The number of overidentifying restriction is 65 in first-differenced model and 76 in the system one. We conduct a sensitivity analysis for changes in the number of instruments and obtain very robust results (see Roodman, 2009).[17]

Table 6: **Estimates for the dynamic log hourly earnings equation with covariates**

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Semikina Wooldridge	*No correction*	*No correction*	*Het. components correction lev eq. only*	*yby correction lev eq. only*	*yby correction all equations*
	GMM	*AB*	*system*	*system*	*system*	*system*
Lag log	0.574***	0.1047**	0.1850***	0.1794***	0.2082***	0.2039***
hourly earnings	(0.040)	(0.0374)	(0.0436)	(0.0442)	(0.0448)	(0.0447)
Education	0.0290***	—	0.0949***	0.0939***	0.0894***	.0931***
	(0.004)		(0.0084)	(0.0083)	(0.0086)	(0.0085)
Age	0.009***	0.0070***	0.0375***	0.03812**	0.02359**	.0228***
	(0.004)	(0.0127)	(0.0113)	(0.0147)	(0.0125)	(0.0126)
Age squared	-0.0001***	-0.0001	-0.0004***	-0.0005***	-0.0002**	-.0003***
	(0.000)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
$\hat{\eta}_i$				0.3020***		
				(0.0833)		
Observations	5033	5033	5033	5033	5033	5033
Joint significance	41.3 (10)	–	–	–	32.91 (11)	14.80 (11)
selection terms	(0.000)				(0.0005)	(0.1920)

Notes.

1. $N = 550$.

2. GMM results obtained using the proposal by Semikyna and Wooldridge (2013).

3. Annual dummies are included in all specifications.

4. *** significant at 1%; ** significant at 5%; * significant at 10%.

5. The standard errors have been corrected following Windmeijer (2005). In columns (4) to (6), we also report corrected standard errors following Terza (2016). See the Appendix for details.

[17]Just for comparison, when we use up to the fourth lag instead of all lags of the log hourly earnings, we obtain the following coefficients: 0.178, 0.093, 0.020 and -0.0002 for the lagged dependent variable, education, age and age squared, respectively. They compare with those in column 3 of Table **6**.

6. The test of significance of the selection terms is a Wald test. Degrees of freedom and level of significance are in parentheses.

The results for the pure autoregressive model are in line with our simulation results. The coefficient of the lagged dependent variable is estimated at 0.103 using the AB estimator and 0.18 using the system GMM estimator without correction. The difference between them may be attributable to the small sample size in the individual dimension.[18] Adding a correction for the correlation of the unobserved heterogeneity components (see column 4), barely changes the coefficient. Alternatively, adding a year-by-year correction in either the equation in levels or in all equations mildly increases the autoregressive parameter. Note, however, that the selection terms are found to be jointly significant.

The autoregressive coefficient (as well as standard errors) remains practically identical in the extended model in Table **6** compared to the pure autoregressive case, and it is substantially lower than the one obtained by SW. The inclusion of the correction terms produces similar results to those in the pure autoregressive model. However, our estimates of the coefficient of the lag of log hourly earnings are in line with the results obtained in a similar context by Arellano *et al.* (1999) using a sample of females from the PSID for the 1970-76 period, and correcting for selectivity (see Table A.3 in that paper). Another dynamic earnings model using the PSID for the 1968-81 period, in this case for males (Holtz-Eakin *et al.*, 1988), yields a similar result for the coefficient of lagged log earnings.

It is also important to note that our age and education estimates are very different from the results in SW, but they are in line with those found in the previous literature using similar data. The coefficients of age, age squared and education have the expected signs, with a quadratic profile of age showing increasing earnings at a decreasing rate. The return to education we get is more in line with the average return to education for females for the US usually found in the literature (see Card, 1999, Harmon *et al.*, 2003 or Polachek, 2008). Regarding endogenous selection, we detect endogenous selection due to correlation between the time-invariant heterogeneity components (column (4) in Table **6**) as well as in column (5), the case in which we have corrected by means of a year-by-year probit the level equations only.

All in all, our opinion is that the similarities among the coefficients with and without correcting for selectivity confirm the results of our Monte Carlo experiment. A lesson for practitioners is that there is little necessity to correct for endogenous selection in situations similar to the one studied in this paper. SW's proposal is only suitable for balanced panels and after making very particular assumptions regarding initial conditions. Although it is feasible to adapt SW's proposal to the more general unbalanced panel case, there are analytical as well as computational costs, which lead us to suggest the simple methods we have just presented in this paper.[19]

[18] An example with large N (4739) small T (6) can be found in Stewart (2007). He presents the results of the estimation of a dynamic panel data model with unbalanced data using GMM methods (Table V). He comments, p. 526, that the AB and system results are substantially identical.

[19] To adapt the SW estimator to an unbalanced panel, we must estimate the model using the SW procedure for each

6 Concluding remarks

In this paper we have analyzed, from the point of view of practitioners, the properties and the performance of GMM-IV estimators of an AR(1) panel data model subject to potentially endogenous sample selection. We show that the Arellano and Bond (1991) and the Anderson and Hsiao (1982) estimators are consistent regardless of the nature (static or dynamic) and the severity of the sample selection process. Alternatively, the Arellano and Bover (1995) and Blundell and Bond (1998) system GMM estimator is moderately biased regardless of the sample size. This implies that to correct the bias induced by endogenous selection, we only need to correct the equations in levels and not the equations in first differences. Note, however, that most of the (small) bias is due to the correlation between the individual heterogeneous components in the outcome and selection equations. All of these results nicely extend to models with exogenous, predetermined or endogenous regressors.

Under these circumstances we evaluate, through a Monte Carlo study of their finite sample properties, the performance of the AB and system GMM estimators in two alternative cases, exogenous (or no selection at all) versus endogenous selection. The results of our experiments confirm the theoretical predictions under a variety of circumstances. We also present an evaluation of some simple alternatives for correction, and showed that univariate corrections based in Wooldridge (1995) are not specially useful for correcting the bias of the system estimator. Alternatively, simple univariate corrections of the potential correlation between the time-invariant heterogeneity components are much more useful. Finally, we do an empirical application confirming available results in the literature of female earnings equations.

In sum, we believe that our theoretical findings, Monte Carlo results, and the empirical application could be of particular relevance for practitioners in cases involving unbalanced data due to sample selection of unknown form (but without feedback from the lagged dependent) or when selection is difficult to model due to missing data problems or lack of appropriate exclusion restrictions.

subpanel (i.e., the subsamples with 4, 5, 6, 7, and so on, observations) and then recover the structural parameters by minimum distance.

References

[1] Anderson, T. W. Hsiao, C. (1982). 'Formulation and estimation of dynamic models using panel data', *Journal of Econometrics*, Vol. 18, pp. 47-82.

[2] Arellano, M. and Bond, S. (1991). 'Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations', *Review of Economic Studies*, Vol. 58, pp. 277-297.

[3] Arellano, M. and Bover, O. (1995). 'Another look at the instrumental-variable estimation of error-components models', *Journal of Econometrics*, Vol. 68, pp. 29-51.

[4] Bover, O. and Arellano M. (1997)'Estimating dynamic limited dependent variable models from panel data', *Investigaciones Economicas*, XXI, 141-165.

[5] Arellano, M., Bover O. and Labeaga, J. M. (1999). 'Autoregressive models with sample selectivity for panel data', in C. Hsiao, K. Lahiri, L. F. Lee, H. Pesaran, H. (eds.), *Analysis of Panels and Limited Dependent Variable Models*, Cambridge University Press, Cambridge, Massachusetts, pp. 23-48.

[6] Baltagi, B. (2005). *Econometric Analysis of Panel Data*, John Wiley and Sons, Chichester.

[7] Blundell, R. and Bond, S. (1998). 'Initial conditions and moment restrictions in dynamic panel data models', *Journal of Econometrics*, Vol. 87, pp. 115-143.

[8] Bover, O. and Arellano, M. (1997). 'Estimating limited-dependent variable models from panel data', *Investigaciones Economicas*, Vol. 21, pp. 141-165.

[9] Card, D. (1999) 'Education and Earnings', In O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*. Amsterdam and New York: North Holland.

[10] Chamberlain, G. (1980). 'Analysis of covariance with qualitative data', *Review of Economic Studies*, Vol. 47, pp. 225–238.

[11] Chamberlain, G. (1984). 'Panel data', in Z. Griliches, M. Intriligator (eds.), *Handbook of Econometrics*, Vol. 2, North-Holland, Amsterdam, Netherlands, pp. 759-798.

[12] Charlier, E., Melenberg, B. and van Soest, A. (2001). 'An analysis of housing expenditures using semiparametric methods and panel data', *Journal of Econometrics*, Vol. 101, pp. 71-107.

[13] Dustman, C. and Rochina-Barrachina, M. E. (2007). 'Selection correction in panel data models: An application to the estimation of females' wage equations', *The Econometrics Journal*, Vol. 10, pp. 263–293.

[14] Fernandez-Val, I. and Vella, F. (2011). 'Bias corrections for two-step fixed effects panel data estimators', *Journal of Econometrics*, Vol. 163, pp. 144–162.

[15] Gayle, G. L. and Viauroux, C. (2007). 'Root-N consistent semiparametric estimators of a dynamic panel-sample-selection model', *Journal of Econometrics*, Vol. 141, pp. 179-212.

[16] González-Chapela J. (2007). 'On the price of recreation goods as a determinant of male labor supply', *Journal of Labor Economics*, Vol. 25, pp. 795-824.

[17] Kazuhiko Hayakawa (2007), "Small sample bias properties of the system GMM estimator in dynamic panel data models", *Economics Letters*, Volume 95, Issue 1, 32-38,

[18] Hansen, L.P. (1982) Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica*, 54(4), 1029-1054.

[19] Harmon, C., Oosterbeek, H. and Walker, I. (2003). 'The returns to education: Microeconomics'. *Journal of Economic Surveys*, Vol. 17, pp. 115-155.

[20] Heckman, James J. "Sample Bias As A Specification Error", Econometrica, 1979, v47(1), 153-162.

[21] Holtz-Eakin, D., Newey, W. and Rosen, H. S. (1988). 'Estimating vector autoregressions with panel data', *Econometrica*, Vol. 56, pp. 1371-1395.

[22] Hu, L. (2002). 'Estimation of a censored dynamic panel data model', *Econometrica*, Vol. 70, pp. 2499-2517.

[23] Jiménez Martín, S. (2006). 'Strike outcomes and wage settlements', *Labour*, Vol. 20, pp. 673-698.

[24] Jiménez Martín, S., Labeaga, J. M. and Rochina-Barrachina, M. E. (2009). 'Comparison of estimators in dynamic panel data sample selection and switching models', Unpublished manuscript.

[25] Jones, A. and Labeaga, J. M. (2004). 'Individual heterogeneity and censoring in panel data estimates of tobacco expenditures', *Journal of Applied Econometrics*, Vol. 18, pp. 157-177.

[26] Kyriazidou, E. (1997). 'Estimation of a panel data sample selection model'. *Econometrica*, Vol. 65, pp. 1335-1364.

[27] Kyriazidou, E. (2001). 'Estimation of dynamic panel data sample selection models'. *Review of Economic Studies*, Vol. 68, pp. 543-572.

[28] Labeaga, J. M. (1999). 'A double-hurdle rational addiction model with heterogeneity: Estimating the demand for tobacco'. *Journal of Econometrics*, Vol. 93, pp. 49-72.

[29] Lai, H. P. and Tsay, W. J. (2016). 'Maximum likelihood estimation of the panel data sample selection model', *Econometric Reviews*, published online.

[30] Mundlak, Y. (1978). 'On the pooling of time series and cross section data', *Econometrica*, Vol. 46, pp. 69-85.

[31] Polachek, S. W. (2008). 'Earnings over the life cycle: The Mincer earnings function and its applications', *Foundations and Trends(R) in Microeconomics*, Vol 4, pp. 165-272.

[32] Raymond, W., Mohnen, P., Palm, F. and van der Loeff S. S. (2010). 'Persistence of innovation in Dutch manufacturing', *The Review of Economics and Statistics*, Vol. 92, pp. 495-504.

[33] Rochina-Barrachina, M. E. (1999). 'A new estimator for panel data sample selection models', *Annales d'Économie et de Statistique*, Vol. 55/56, pp. 153-181.

[34] Roodman, D. (2009). 'A note on the theme of too many instruments', *Oxford Bulletin of Economics and Statistics*, Vol 71, pp. 135-158.

[35] Roodman, D. 2006. How to Do xtabond2: An introduction to "Difference" and "System" GMM in Stata. Working Paper 103, Center for Global Development, Washington.

[36] Sasaki, Y. (2015). 'Heterogeneity and selection in dynamic panel data', *Journal of Econometrics*, Vol. 188, pp. 236-249.

[37] Semykina, A. and Wooldrigde J. M. (2010). 'Estimating panel data models in the presence of endogeneity and selection: Theory and application', *Journal of Econometrics*, Vol. 157, pp. 375-380.

[38] Semykina, A. and Wooldrigde, J.M. (2013). 'Estimation of dynamic panel data models with sample selection', *Journal of Applied Econometrics*, Vol. 28, pp. 47-61.

[39] Semykina, A. and Wooldrigde, J.M. (2018). Binary response panel data models with sample selection and self-selection', *Journal of Applied Econometrics*, Vol. 33, pp. 179-197.

[40] Stewart, M. (2007) 'The interrelated dynamics of unemployment and low-wage employment', *Journal of Applied Econometrics*, Vol. 22, pp. 511-531.

[41] Terza, J. V. (2016) 'Simpler standard errors for two-stage optimization methods', *The Stata Journal*, Vol. 16, pp. 368-385.

[42] Vella, F. and Verbeek, M. (1998). 'Two-step estimation of panel data models with censored endogenous variables and selection bias', *Journal of Econometrics*, Vol. 90, pp. 239-263.

[43] Verbeek, M. and Nijman, T. (1992). 'Testing for selectivity bias in panel data models', *International Economic Review*, Vol. 33, pp. 681-703.

[44] Winder, K. L. (2004). 'Reconsidering the motherhood wage penalty', Unpublished manuscript.

[45] Windmeijer, Frank, (2005), A finite sample correction for the variance of linear efficient two-step GMM estimators, Journal of Econometrics, 126, issue 1, p. 25-51.

[46] Wooldrigde, J.M. (1995). 'Selection corrections for panel data under conditional mean independence assumptions', *Journal of Econometrics*, Vol. 68, pp. 115-132.

[47] Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data (2nd ed.). Cambridge, Mass.: MIT Press.

Appendix

A Appendix to the consistency of the estimators

Consider the linear model

$$y = Y'\theta + u,$$

where Y is endogenous and y is a response scalar variable. We assume that we have an exogenous set of instruments z. Define

$$u(\theta) = y - Y'\theta.$$

The sample selection process is given by $s = s_z s_y s_Y$, i.e. a data point (y, Y, z) is available if and only if all three variables are available. The classical condition for exogeneity is that

$$E(u(\theta_0)|s, z) = 0.$$

See p. 795 of Wooldridge (2010). However, this condition can be difficult to verify in some contexts, particularly in a dynamic panel setting such as the case presented in this paper. The alternative condition

$$E(s_y s_Y u(\theta_0)|s_z, z) = 0$$

can be much easier to verify and still leads to consistency. Recall that under the usual conditions, the consistency of the GMM estimator of θ requires that $E(szu(\theta)) = 0$ if and only if $\theta = \theta_0$. This is easily proven,

$$E(szu(\theta_0)) = E(s_z z s_y s_Y u(\theta_0)) = E(s_z z E(s_y s_Y u(\theta_0)|s_z, z)) = 0$$

On the other hand, for $\theta \neq \theta_0$,

$$E(szu(\theta)) = E(szu(\theta \pm \theta_0)) = E(szu(\theta_0)) - E(szY')(\theta - \theta_0) = E(szY')(\theta_0 - \theta).$$

Therefore, it suffices to have $\text{rank}(E(szY')) = \dim(\theta)$, which is to say the instruments have a full effect on the endogenous variables in the observed sample.

B The variance of corrected estimators

Assume that the relationship among variables, instruments and parameters (for $l = 1, \ldots, L$ moments) is given by the following expression:

$$m_l(y_i, x_i, z_i, \theta) = \frac{1}{N}\sum_{i=1}^{N} m_{il}(y_i, x_i, z_i, \theta) = \frac{1}{N}\sum_{i=1}^{N} m_{il}(\theta)$$

Then, we can define the objective function, for instance, as:

$$q = \sum_{l=1}^{L} m_l^2$$

with

$$m_l = \frac{1}{N}\sum_{i=1}^{N} m_{il}(\theta) = 0$$

We choose θ which minimises:

$$q = m(\theta)' A m(\theta)$$

with A being any semi-definite positive matrix, which is not a function of θ. We can choose the asymptotic variance of $m(.)$, say W, so that the estimator solving the problem:

$$q = m(\theta)' W^{-1} m(\theta)$$

is the GMM estimator. The best option for the variance-covariance matrix of the GMM estimator, as suggested by Hansen (1982), is:

$$V_{GMM} = \left[G' W^{-1} G \right]^{-1}$$

where G is a matrix of derivatives whose j row is:

$$G^{jl} = \frac{\partial m_l(\theta)}{\partial \theta'}$$

Because the criterion is linear in θ, the solution for $\hat{\theta}$ can be expressed linearly, and its variance-covariance matrix is $\left[X' Z_1 \hat{W} Z_1' X \right]$, where Z_1 is the matrix of instruments, and all matrices should be defined conditional on the selected sample. The optimal choice for $\hat{W}$ is $\left[Z' \hat{u} \hat{u}' Z \right]$. Because we estimated in a first step $\hat{\lambda}_{it}(z_{it}\hat{\gamma})$, using univariate probits for each T, we must correct $\hat{W}$ to take that into account. We can do this correction using the scores of the likelihood function for this parameter evaluated at the optimal maximum likelihood estimates. If Z is the matrix of exogenous regressors used to adjust the probit model, we can use for the correction, for instance, $Z' C Z$, with

$$C = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial l_{it} \partial l_{is}}{\partial \lambda_t \partial \lambda_s'}$$

where l_{it} is the likelihood function for individual i in period t. A simpler alternative to calculate the estimated asymptotically correct covariance matrix of the first-differenced GMM-IV and system GMM-IV estimators after correcting for sample selection, which we used here according to Terza (2016). It involves the scores of the likelihood function at each period, but there is no need to calculate C.

Table A1. Average moment conditions of simulated errors and most recent instruments

$N = 500$	$E(\Delta\varepsilon_{it}y_{it-2}/A_{it})$	$E((\alpha_i + \varepsilon_{it})\Delta y_{it-1}/A_{it})$	$E(\varepsilon_{it}\Delta y_{it-1}/A_{it})$	$E(\alpha_i\Delta y_{it-1}/A_{it})$
$corr(\varepsilon_{it}, u_{it}) = 0.242 = corr(\alpha_i, \eta_i)$				
$\rho = 0.25$	-.0021	.0020	.0015	.0004
$\rho = 0.50$	-.0036	.0008	.0017	-.0009
$\rho = 0.75$	-.0071	.0001	.0019	-.0018
$corr(\varepsilon_{it}, u_{it}) = 0.242$; $corr(\alpha_i, \eta_i) = 0$				
$\rho = 0.25$	-.0021	.0020	.0015	.0005
$\rho = 0.50$	-.0037	.0018	.0017	.0002
$\rho = 0.75$	-.0071	.0020	.0019	.0001
$corr(\varepsilon_{it}, u_{it}) = 0.447 = corr(\alpha_i, \eta_i)$				
$\rho = 0.25$	-.0011	.0012	.0019	-.0007
$\rho = 0.50$	-.0025	-.0014	.0030	-.0044*
$\rho = 0.75$	-.0057	-.0037	.0042**	-.0079***
$corr(\varepsilon_{it}, u_{it}) = 0.447$; $corr(\alpha_i, \eta_i) = 0$				
$\rho = 0.25$	-.0011	.0025	.0019	.0006
$\rho = 0.50$	-.0026	.0031	.0030	.0001
$\rho = 0.75$	-.0057	.0042	.0042**	-.0000
$N = 5000$	$E(\Delta\varepsilon_{it}y_{it-2}/A_{it})$	$E((\alpha_i + \varepsilon_{it})\Delta y_{it-1}/A_{it})$	$E(\varepsilon_{it}\Delta y_{it-1}/A_{it})$	$E(\alpha_i\Delta y_{it-1}/A_{it})$
$corr(\varepsilon_{it}, u_{it}) = 0.242 = corr(\alpha_i, \eta_i)$				
$\rho = 0.25$	.0016	-.0001	-.0001	-.0015***
$\rho = 0.50$	.0019	-.0019**	.0003	-.0022***
$\rho = 0.75$	.0035	-.0022**	.0008	-.0030***
$corr(\varepsilon_{it}, u_{it}) = 0.242$; $corr(\alpha_i, \eta_i) = 0$				
$\rho = 0.25$	.0015	-.0009	-.0001	-.0008
$\rho = 0.50$	.0019	-.0006	-.0003	-.0009
$\rho = 0.75$	.0034	-.0002	.0008	-.0009*
$corr(\varepsilon_{it}, u_{it}) = 0.447 = corr(\alpha_i, \eta_i)$				
$\rho = 0.25$	.0017	-.0019*	.0014*	-.0033***
$\rho = 0.50$	.0022	-.0035***	.0027***	-.0062***
$\rho = 0.75$	.0044	-.0051***	.0041***	-.0091***
$corr(\varepsilon_{it}, u_{it}) = 0.447$; $corr(\alpha_i, \eta_i) = 0$				
$\rho = 0.25$	.0016	.0005	.0014*	-.0008
$\rho = 0.50$	.0020	.0017*	.0027***	-.0010
$\rho = 0.75$	.0041	.0030***	.0041***	-.0011*

Notes.

1. 1000 simulations.

2. Static selection model (A).

3. $A_{it} = z_{it}, d_{it} = d_{it-1} = d_{it-2} = 1$.

4. *** significant at 1%; ** significant at 5%; * significant at 10%.