Consistent estimation of panel data sample selection models

Badi H. Baltagi

Department of Economics and Center for Policy Research

Syracuse University

426 Eggers Hall, Syracuse, NY 13244-1020, US

Sergi Jiménez-Martín*

Department of Economics and BSE

Universitat Pompeu Fabra

Ramon Trias Fargas, 25-27, 08005-Barcelona, Spain

José M. Labeaga

Department of Economics

UNED

Senda del Rey s/n, 28040-Madrid, Spain

Majid al Sadoon

Business School

Durham University

Mill Hill Lane, Durham, DH1 3LB, UK

November 2023

Abstract

The properties of classical panel data estimators including fixed effect, first-differences, random effects, and generalized method of moments-instrumental variables estimators in both static as well as dynamic panel data models are investigated under sample selection. The correlation of the unobserved errors is shown not to be sufficient for the inconsistency of these estimators. A necessary condition for this to arise is the presence of common (and/or non-independent) non-deterministic covariates in the selection and outcome equations. When both equations do not have covariates in common and independent of each other, the fixed effects, and random effects estimators in static models with exogenous covariates are consistent. Furthermore, the first-differenced generalized method of moments estimator uncorrected for sample selection as

*Corresponding author: Sergi Jiménez-Martín. sergi.jimenez@upf.edu.

well as the instrumental variables estimator uncorrected for sample selection are both consistent for autoregressive models even with endogenous covariates. The same results hold when both equations have no covariates in common but are correlated once we account for such correlation. Under the same circumstances, the system generalized method of moments estimator adding more moments from the levels equation has moderate bias. Alternatively, when both equations have common covariates the appropriate correction method is suggested. Serial correlation of the errors being a key determinant for that choice. The finite sample properties of the proposed estimators are evaluated using a Monte Carlo study. Two empirical illustrations are provided.

JEL Codes: J52, C23, C24

Keywords: Panel data, sample selection, generalized method of moments, fixed and random effects, differenced estimator

1 Introduction

The problems of self-selection, non-response and attrition are common in datasets containing economic variables. Their presence is well researched in cross-section studies. However, correlated heterogeneity together with endogenous attrition, non-response or sample selection complicate matters with unbalanced panel data (Baltagi, 2021). The increasing availability of large longitudinal databases has produced many studies simultaneously dealing with unobserved heterogeneity and selectivity. Moreover, the development of new methods make these approaches more likely to be used in the future. In this context, we believe that it is important to highlight the advantages and disadvantages of various commonly used panel data estimators and to draw the researchers' attention to potential pitfalls in using them in empirical studies.

In this paper we focus on the estimation of a very general class of panel data sample selection models. We consider a variety of cases for the outcome of interest and a simple form for the selection equation. We also allow for a very general correlation structure in the error components of both equations. Departing from the simplest situation, we present an exercise which includes some important features in the model to test their individual and joint effects on the bias of some of the classical estimators (fixed effects -FE-, random effects -RE-, and first differences -FD-) as well as the generalized method of moments (GMM) estimators.

In more detail, we consider four cases of increasing complexity: (a) panel data sample selection models without common covariates, and independent on each other; (b) models without common covariates, but dependent on each other; (c) models with at least one common covariate but not serially cross-correlated time-variant errors; and, (d) models with at least one common covariate and time variant serially cross-correlated errors.

The first two cases are less common than others. They typically involve sample selection related to involuntary factors, not linked to the individual characteristics (the Covid-19 crisis or the increasingly common empirical studies based on an experimental or quasi-experimental designs, for example). In this context, the determinants at the extensive margin are completely different from those at the intensive margin. Some examples of standard economic models imposing identification restrictions that exclude from the observability rule variables included in the outcome equation include Rochina-Barrachina (1999), or Knoef and Been (2015). Under these assumptions, sam-

ple selection corrections are not necessary for consistent estimation of the parameters of interest. However, sample selection corrections (*a la Heckman*) are necessary in the last two more common cases. Finally, correlation between the unobserved components can also cause endogenous sample selection.

For cases (a) and (b) we distinguish between static and dynamic sample selection models. In the static model without common (and independent) covariates between the outcome and the selection equations (let us call them x and z respectively), we show that the classical panel data estimators (FE, FD, RE GLS and GMM) are all consistent. Similarly, in dynamic models without common time-varying covariates such as the purely AR(1) as well as the Monte Carlo study in Raymond et al. (2007), the GMM estimator proposed by Arellano and Bond (AB, 1991) as well as the less efficient Anderson and Hsiao (AH, 1982) estimators, both uncorrected for sample selection, are consistent regardless of the exogenous or endogenous nature of the selection. An immediate implication of this result is that GMM estimators are not consistent in the uncorrected model when the lagged outcome is part of the selection equation.

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 bias of the system GMM estimator is small especially when the time-invariant heterogeneous components in the outcome and selection equations are not correlated. This also applies to models with exogenous, predetermined or endogenous covariates, which are, in turn, not present in the selection equation.

For models with at least one common covariate (case (c)), which could be the lagged outcome as in Gayle and Viaurous (2007), we unify and extend some of the most popular approaches. In particular, we propose an extension of Wooldridge (1995) and Rochina-Barrachina (1999) based on either the simple estimation of year-by-year probits or the adjustment of bivariate probits to build the corrections since our model, contrary to Wooldridge (1995) and Rochina-Barrachina (1999), is dynamic and the selection imposes the condition on three consecutive positive events to have a usable observation. In static models in levels, we follow Wooldridge (1995) and correct for selection bias by adding the current selection term. In first-differenced models and, in general, in dynamic models, the complexity of the correction critically depends on the serial correlation of the errors. In the simplest case (no serial correlation and stationarity) we show that the Wooldridge's proposal

can be applied and, more importantly, extended to dynamic models with the necessary adjustments.

Finally, when both equations have common covariates and the time varying errors are serially cross-correlated (case (d)), we suggest, following Rochina-Barrachina (1999), a multivariate correction adapted to the dynamic case. In models with predetermined or endogenous covariates the selection terms need to be instrumented accordingly.

Testing between the alternative cases described above is not complicated. For example, a simple t-test or Wald test allows checking for the significance of x in the selection equation. In case it is not detected, a test of the $E(x|z)$ checks for the need to correct for the correlation between x and z. Finally, to distinguish between (c) and (d) we can test the correlation between the time-varying errors in the outcome and the lagged (once and twice if necessary) time-varying errors in the selection equations.

The performance of these estimators is evaluated using Monte Carlo methods, relaxing or imposing a variety of assumptions. In models without common covariates in both equations, our results suggest that there is no need for correcting the classical panel data static estimators or the first-differences dynamic panel AB estimates in the selected sample. In models with common covariates, we show that our suggested estimator is able to control for selection bias. This paper highlights the advantages and disadvantages of various methods. This should prove useful for applied work in this area.

Our work contributes to the literature in several dimensions. First, it shows that it is unnecessary to correct for selectivity (even with a high degree of correlation) when both equations do not have common time-varying covariates. Second, it suggests simple methods to correct the outcome equation when both equations have common covariates. Combining these contributions, we conclude that a key factor of the necessity of sample selection correction *a la Heckman* is the presence of common covariates in both equations along with correlation of the errors and not whether the errors of both equations are correlated alone. 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, when exclusion restrictions are not available, or in experimental or quasi-experimental settings where the selection and outcome equations contain different sets of determinants.

The outline of the paper is as follows: Section 2 presents a general framework and the estimation

strategies. Section 3 shows under what conditions many standard panel data estimators remain consistent under sample selection. The performance of the proposed estimators is studied in Section 4. We present Monte Carlo results showing the finite sample average bias in many relevant cases. In Section 5, we present two empirical applications to illustrate several features or our theoretical and simulation results. Section 6 concludes.

2 A general framework

In this section, we consider a flexible framework that nests all the cases we study. We focus a dynamic panel data model with unobserved heterogeneity. We must note that the first proposals appeared in a static context (see Verbeek and Nijman, 1992, Wooldridge, 1995, Rochina-Barrachina, 1999 and Kyriazidou, 1997, for models with strict exogeneity and Vella and Verbeek, 1998, and Semykina and Wooldridge, 2010 allowing for endogenous explanatory variables). In another strand of research, theoretical papers have explored bias-corrected estimators for the static case (Fernández-Val and Vella, 2011). More recently, Sasaki (2015) discussed non-parametric identification of panel data selection models and Lai and Tsay (2018) proposed maximum simulated likelihood methods in a static set-up. In particular, we consider the following model:

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

for $i = 1, ..., N$ and $t = 1, ..., T$, where y^* is the latent outcome, which is observed when d^*, the observability criteria (defined below) is greater than zero. Furthermore x is a vector of covariates which, for ease of exposition, we simplify as a single covariate, that can be either exogenous, predetermined or endogenous, and α_i is an individual heterogeneous component independent of the idiosyncratic error ε_{it} but potentially correlated with x. A model like (1) appeared for the first time in Arellano *et al.* (1999) and Kyriazidou (2001). More recently, Semykina and Wooldridge (2013) introduced new two-stage random effects strategies for estimating panel data models in the presence of endogeneity, dynamics and selection.

Different values of ρ and β lead to different models. For example, $\rho = 0$ leads to a static panel data model; $|\rho| < 1$ and $\beta = 0$ yield a purely stationary AR(1); of course, when both parameters are different from zero we have an autoregressive model with covariates.

We assume the following process for x:

$$x_{it} = \rho_x x_{it-1} + \phi_x \wp_{it} + \alpha_i^x + \varepsilon_{it}^x, \tag{2}$$

where $-1 < \rho_x < 1$, $\wp$ is a strictly observed exogenous covariate, α_i^x is a heterogeneity component and ε_{it}^x is a time-variant error component. Note that the process for x can be easily generalized without affecting any fundamental result in this paper. In case x is exogenous both error components are uncorrelated with other errors components in the model; when x is predetermined we allow correlation with ε_{it-1}; and finally, when x is endogenous we allow correlation between the error components in (1) and (2).

In the case of selection, the variable of interest is partially observed, and it is usual to specify an observability or selection rule of the form:

$$d_{it}^* = z_{it}\gamma + \delta x_{it} + \eta_i + u_{it}, \tag{3}$$

where η_i is a term capturing unobserved individual heterogeneity that can be correlated with both z and x, z_{it} is a vector of strictly exogenous regressors including a constant and x_{it} is the same (vector of) regressor(s) that appears also in the outcome equation. Our framework also allows the case where x is the lagged outcome y_{t-1}. While this makes identification more difficult, it fits well in our general argument. Regarding the correlation structure of the covariates, we assume that z and x do not have variables in common and so, z represents exclusion restrictions. For ease of exposition, we assume that z and x are not correlated with η_i. However, none of the main results of the paper are affected in case we allow correlation of z or x with η_i or x with α_i (as we show in Appendix B, if there is correlation, we can express $\eta_i = g(z_i, x_i)$ and add this function as additional regressors following either Mundlak, 1978, or Chamberlain, 1984). Finally, u_{it} is a time varying error. The observed indicator d_{it} is given by:

$$d_{it} = 1[d_{it}^* > 0] = 1[z_{it}\gamma + \delta x_{it} + \eta_i + u_{it} > 0], \tag{4}$$

such that $d_{it} = 1$ if $y_{it}^* = y_{it}$, when the latent outcome is observed, and zero otherwise.

The error components in equation (1) are related to the error components in the selection

equation as follows:

$$\alpha_i = \alpha_i^0 + \theta_0 \eta_i, \tag{5}$$

and

$$\varepsilon_{it} = \varepsilon_{it}^0 + \vartheta_0 u_{it} + \vartheta_1 u_{it-1} + \vartheta_2 u_{it-2}, \tag{6}$$

where, for simplicity, α_i^0 and ε_{it}^0 are assumed to be normally distributed and θ_0 and $\vartheta_j; j = 0, 1, 2$ are the parameters introducing correlation. In case they are all zero, there is exogenous sample selection. Alternatively, when any of them is different from zero, there is endogenous sample selection. We distinguish between two cases: A) the contemporaneous correlation case, when $\vartheta_0 \neq 0$ and $\vartheta_j = 0; j = 1, 2$; and, B) the more complex case of serial cross-correlation, when $\vartheta_j \neq 0; j = 0, 1, 2$.

It is well known that in the absence of endogenous selection and for the typical situation of N large and T small, the outcome equation can be estimated with standard methods. In the static case, $(\rho = 0)$, with exogenous regressors, FE and RE estimators are consistent under the additional assumption that α_i and x are not correlated. In case x and α_i are correlated, i.e., $\alpha_i = g(x_i) + \alpha_i^*$ where α_i^* is an error term independent of x_i, we can, following Wooldridge, add to (1) the control function $g(x_i)$. Alternatively, for the purely AR(1) or the dynamic model with covariates, these are consistently estimated using IV methods including Anderson and Hsiao (1982), Arellano and Bond (1991), Arellano and Bover (1995) and Blundell and Bond (1998).

2.1 Estimation of the model

2.1.1 The static case, $\rho = 0$

Estimation in levels: Equation (1) could be estimated in levels by RE. In the case x_{it} is strictly exogenous, a sufficient condition for the RE estimator to yield consistent estimates is the following:

$$E(\alpha_i + \varepsilon_{it} | x_{it}, d_{it} = 1) = E(\alpha_i | x_{it}, d_{it} = 1) + E(\varepsilon_{it} | x_{it}, d_{it} = 1) = 0 \quad \forall t. \tag{7}$$

As a general rule, RE estimates on the selected subsample are inconsistent if selection is non-random, and/or if there is correlated individual heterogeneity.

Estimation in time differences: Again, if x_{it} is strictly exogenous, a sufficient condition for the differenced estimator to be consistent is the following:

$$E(\varepsilon_{it} - \varepsilon_{it-s}|x_{it}, x_{is}, d_{it} = d_{is} = 1) = 0 \; s < t. \tag{8}$$

If condition (8) is satisfied, the differenced estimator (also the within-groups estimator which also wipes out the individual effects) provide consistent estimates. Alternatively, if this condition is violated consistent estimation requires considering the selection process. In this sense, Dustmann and Rochina-Barrachina (2007) compare the methods proposed by Wooldridge (1995), Kyriazidou (1999) and Rochina-Barrachina (1999) in the estimation of static females' wage equations.

2.1.2 The AR(1) and dynamic cases

In the small T dynamic case, IV methods are in general necessary (as is well-known, when T is sufficiently large, we can consistently estimate the parameters of the model using the within-groups estimator, see Nickell, 1991). As pointed out above, we consider the following estimation options: 2SLS-IV (AH: Anderson & Hsiao, 1982) and, more generally, GMM (AB: Arellano & Bond, 1991; System GMM: Arellano & Bover, 1995; Blundell & Bond, 1998). All of these estimators require first differencing the data (and using also the equations in levels in the case of the system GMM estimator). They also use internal instruments lagged at least twice, which implies that the selected sample is conditional on observing the outcome for at least three consecutive periods ($d_{it}, d_{it-1}, d_{it-2} = 1$). Although the AH and AB estimators are two well-known methods, the system GMM ones deserves further explanations, first, because it is not as common in empirical applications and, second, to relate the four IV methods used.

Arellano and Bond (1991) propose a dynamic panel data estimator that generalizes the Anderson and Hsiao (1981) estimator by using more orthogonality conditions that exist between the lagged values of the dependent variable and the error component disturbances. Both estimators difference the model to eliminate the unobserved heterogeneity, see Baltagi (2021, pp.189-191) for details.

Arellano and Bover (1995) stack a system of equations, one averaged over time and hence a levels equation, on top of a forward orthogonalized equation eliminating the individual error component and generalize the Hausman and Taylor (1981) estimator to obtain an efficient GMM estimator

of a dynamic panel data model using more moments than the Arellano and Bond estimator, see Baltagi (2021, pp.194-198) for details.

Blundell and Bond (1998) exploit an additional mild stationarity restriction on the initial conditions to generate a system GMM estimator that uses more moment conditions than Arellano and Bond (1991). Essentially, they use lagged levels of the dependent variable as instruments for the equation in first differences as in Arellano and Bond (1991). Additionally, the stationarity restriction on the initial condition allows the use of lagged differences of the dependent variable as instruments for an equation in levels, see Baltagi (2021, pp. 201-203) for details.

For the AH and the AB to be consistent, we need the following orthogonality condition to hold:

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

which is stronger than the orthogonality condition imposed in the standard case. Note that when this restriction holds, it also holds for $t-3$ and backward lags. For the consistency of the system GMM estimator, we need the following condition:

$$E[(\alpha_i + \varepsilon_{it})\Delta y_{it-1}|d_{it} = d_{it-1} = d_{it-2} = 1)] = 0, \tag{10}$$

which is also stronger than the orthogonality condition imposed in the standard case.

Arellano *et al.* (1999) proposed the estimation of sample selection models conditioning on exogenous positive past outcomes for at least three consecutive previous periods and showed that the degree of selection is significantly reduced in economic models with persistence.

2.2 Estimation under endogenous sample selection

In the presence of endogenous sample selection in the standard static case, researchers usually proceed using the method developed by Wooldridge (1995). It is worth mentioning that his estimator for static linear unobserved components panel data models allows correlation between the unobserved component (FE) and observable explanatory variables, without imposing distributional assumptions on the unobserved effect. The idiosyncratic errors in the regression equation can have serial dependence of unspecified form. Wooldridge's (1995) estimator goes a step further than previous methods, which considered RE under the assumptions of normality and serial independence

of the idiosyncratic errors in both the selection and regression equations, and the time-constant unobserved effects in the selection and regression equations. The latter are assumed to be normally distributed (see Verbeek and Nijman, 1992). First, one corrects the problem of endogenous selection induced by the correlation of the errors in both equations, and, then, one estimates the outcome equation. Since, contrary to Wooldridge (1995), we also propose dynamic models, we need to distinguish between two cases:

A. When there is some feedback between the (time variant non-deterministic) covariates (when the common covariates are deterministic or time-invariant there is no need to correct estimates in first-differences and, as we will see later on, little necessity to correct estimates in levels), in the outcome and the selection equation. The need for sample selection correction varies with the sampling condition and the correlation structure of the errors in both equations. We consider two cases:

   A1. Contemporaneous correlation: $\vartheta_0 \neq 0$ and $\vartheta_j = 0; j = 1, 2;$

   - Step 1. Following Wooldridge (1995), we estimate year-by-year probit models and compute univariate correction terms (Heckman's lambda).
   - Step 2. Add the appropriate selection terms as additional regressor(s) to the relevant outcome equation. In Appendix B we show that when the errors are not serially correlated, univariate corrections are sufficient regardless of the observability condition: one observation in static level models (see equation (A12) in Appendix B), two and three consecutive observations in, respectively, first-differenced static models (see Rochina-Barrachina, 1999) or dynamic models (equation (A9) in Appendix B). We estimate the equation of interest including the appropriate correction(s) using one of the methods described in Table **1**.

     For example, in the case of a pure AR(1) model, the sample has to be selected in three consecutive periods to have a usable observation in the current period. Then, the appropriate correction involves the current lambda in the equation in levels and the first-differenced lambda in the first-differenced equation (see Jiménez-Martín, 1999, 2006). Under contemporaneous correlation, standard software can be used (see, for instance, Roodman, 2006). Corrected standard errors need to be computed

anyway. This can be done by means of the delta method or bootstrapping.

We must also note that Rochina-Barrachina (1999), in the context of a static model, proposed an estimator that relaxes some of the assumptions in the Wooldridge (1995) method. Specifically, the estimator allows for an unknown conditional mean of the individual effects in the main equation. This allows the use of an alternative set of identifying restrictions to overcome the selection problem. In particular, the estimator imposes that the joint distribution of the time differenced regression equation error and the two selection equation errors, conditional upon the entire vector of (strictly) exogenous variables, is normal.

A2. Cross serial correlation: $\vartheta_j \neq 0; j = 0, 1, 2;$

- Step 1. When the correlation structure of the errors is complex, a more sophisticated bivariate or trivariate correction is required, either in static models with endogenous regressors or in dynamic models. Following Rochina-Barrachina (1999) and Jiménez-Martín *et al.* (2009), we propose estimating bivariate and trivariate probit models of, respectively, the probability that $d_{it} = d_{it-1} = 1$ and $d_{it} = d_{it-1} = d_{it-2} = 1$ (see Appendix B).
- Step 2. Under stationary correlation and exchangeability (Kiriadizou, 1997), the first-differenced equations require two correction terms obtained, under normality, from the previous estimated trivariate probit model (equation (A8) in Appendix B). Alternatively, the equation in levels requires also two correction terms but, in this case, obtained from a bivariate probit (equation (A11) in Appendix B). Note that, since the equations in first differences and levels require different corrections, we suggest either using the Stata ***gmm*** routine.

B. When there is no feedback between the outcome and the selection equations, i.e., when $x \perp z$ and x is not part of the selection equation. This is the case of the purely AR(1) model as well as models of attrition or missing variables where the reason for selecting the sample is correlated with the object of study but unrelated to other determinants of the model. These assumptions are not going to be maintained in labor supply models, wage equations, etc. In this context the following results hold:

- Result 1: Under endogenous selection and absence of feedback from the outcome equation to the selection equation it is feasible to show that the AH and the AB estimators are both consistent. In fact, for the AB estimator

$$E[\Delta\epsilon_{it}y_{it-k}|d_{it}, d_{it-1}, d_{it-2} = 1] = 0 \quad k > 1,$$

and, for the AH estimator

$$E[\Delta\epsilon_{it}y_{it-2}|d_{it}, d_{it-1}, d_{it-2} = 1] = 0.$$

Furthermore, the AH and the AB estimators are consistent (with the same asymptotic distribution as the original AH and AB estimators) in the model with either exogenous, predetermined or endogenous covariates.

An implication of Result 1 is that it applies to the case in which a deterministic or time-invariant covariate x is included in the selection equation.

- Result 2: Under the same conditions above (correlation of the time-variant and time-invariant error components) the system GMM estimator is not consistent since

$$E[\epsilon_{it}\Delta y_{it-1}|d_{it}, d_{it-1}, d_{it-2} = 1] \neq 0.$$

However, our Monte Carlo results show that the bias is small, especially when the individual heterogeneous components are not correlated. Moreover, in the model with covariates, the system GMM estimator has a small bias under the same conditions, regardless of the nature of the covariates.

Follow-up to result 2: To correct the bias of the system GMM estimator, we need to correct for selection only in the levels equation. If the correlation between the time-invariant error components is zero and there is no feedback between both equations, the bias of the system GMM estimator is small (but not zero). So, when the AB estimator does not work well (small N, large autoregressive coefficient), the system GMM estimator is highly recommended.

- Result 3: The previous results can be extended to static panel data models regardless of

the nature of the covariates. This implies that, when there is no feedback between the outcome and the selection equations ($x \perp z$ and x is not part of the selection equation), we can recover consistent estimates using either FE, FD or RE (GLS) methods (consistency of the GLS estimator requires a preliminary step in order to account for the possibility that $cov(x_i, \alpha_i) \neq 0$)).

- Result 4: When x is not present in the selection equation but is not independent from z it is still possible to avoid bias correction *a la Heckman* by accounting for the relation between x and z, $E(x|z)$ say, in the outcome equation.

In Table **1** we summarize all the cases considered and the suggested solutions. We distinguish between four static and five dynamic models. As we show in the next section, when there are no common covariates between both equations and they are independent, there is no need to correct for sample selection for the static estimators and some of the dynamic ones (AH and AB). In case they are not independent, a control function approach (based on the $E(x|z)$) can account for any potential bias induced by the selection process. Alternatively, when at least a time-varying covariate is included in both equations sample selection corrections (either univariate or multivariate, depending on the serial cross-correlation of the errors) are required to get consistent estimates.

Table 1: Models considered under endogenous sample selection: Cases and solutions[1]

Model	AR param	x in outcome	x endog	x in selection	Correction needed	Estimation methods
Static	$\rho = 0$	Yes	No	No	No	FE, RE(GLS)[2], FD
Static	$\rho = 0$	Yes	Yes	No	No	FD-IV, FD-GMM
Static	$\rho = 0$	Yes	No	Yes	Yes	FE, RE(GLS)[2], FD
Static	$\rho = 0$	Yes	Yes	Yes	Yes	FD-IV, FD-GMM
AR(1)	$\lvert\rho\rvert < 1$	No	—	nr	No	FD-IV, FD-GMM
Dynamic	$\lvert\rho\rvert < 1$	Yes	No	No	No	FD-IV, FD-GMM
Dynamic	$\lvert\rho\rvert < 1$	Yes	Yes	No	No	FD-IV, FD-GMM
Dynamic	$\lvert\rho\rvert < 1$	Yes	No	Yes	Yes	FD-IV, FD-GMM
Dynamic	$\lvert\rho\rvert < 1$	Yes	Yes	Yes	Yes	FD-IV, FD-GMM

Notes.
1. We assume $x \perp z$. When this assumption does not hold and x is not present in the selection equation we will follow a control function approach to consider this correlation.
2. Consistency of the GLS estimator relies strongly on the assumption that $cov(x_i, \alpha_i) = 0$. When this assumption does not hold we follow either Chamberlain (1984) or Mundlak (1978) approach to account for the correlation between x and α or η as well as the correlation of z and η.

3 Consistency under endogenous sample selection

In this section we analyze the consistency of potential estimators as a function of a key factor: the presence of common time-varying covariates in the outcome and selection equations. We show that many standard estimators are consistent regardless of the correlation between the errors in the selection and the outcome equations when there are no common covariates between them. For example, for dynamic models the AH and AB estimators are consistent when the outcome and selection equations have no regressors in common, i.e., when all the regressors in the selection equation are exclusion restrictions. The system GMM estimator is an exception and has a small bias, mainly induced by the correlation between the time-invariant heterogeneous components in the outcome and the selection equation.

3.1 Consistency in the pure autoregressive model

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

$$y^*_{it} = \alpha_i + \rho_0 y^*_{it-1} + \varepsilon_{it}, \tag{11}$$

$$d_{it} = 1(\eta_i + \gamma_0 z_{it} + u_{it} > 0), \tag{12}$$

$$\alpha_i = \alpha^0_i + \theta_0 \eta_i, \tag{13}$$

and

$$\varepsilon_{it} = \varepsilon^0_{it} + \vartheta_0 u_{it}. \tag{14}$$

The exogenous random variables 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. We assume that $E(\varepsilon^0_{it}) = E(u_{it}) = 0$. 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 show $E(s_{ist} y^*_{is}\Delta\varepsilon_{it}(\rho)) = 0$ iff $\rho = \rho_0$, where s_{ist} is defined as

$$s_{ist} = d_{it} d_{it-1} d_{it-2} d_{is}. \tag{15}$$

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

$$\begin{aligned} E(s_{ist}y^*_{is}\Delta\varepsilon_{it}(\rho)) &= E(s_{ist}y^*_{is}(\Delta y^*_{it} - \rho\Delta y^*_{it-1})) \\ &= E(s_{ist}y^*_{is}(\rho_0\Delta y^*_{it-1} + \Delta\varepsilon_{it} - \rho\Delta y^*_{it-1})) \\ &= (\rho_0 - \rho)E(s_{ist}y^*_{is}\Delta y^*_{it-1}) + E(s_{ist}y^*_{is}\Delta\varepsilon_{it}). \end{aligned} \tag{16}$$

Identification requires that $E(s_{ist}y^*_{is}\Delta y^*_{it-1}) \neq 0$ and $E(s_{ist}y^*_{is}\Delta\varepsilon_{it}) = 0$. A classic sufficient condition that ensures exogeneity is $E(\Delta\varepsilon_{it}|s_{ist}, y^*_{is}) = 0$. However, it is not feasible to verify this condition in our context. A simpler sufficient condition derived in the Appendix A is the following

$$E(d_{it}d_{it-1}d_{it-2}\Delta\varepsilon_{it}|d_{is}, y^*_{is}) = 0. \tag{17}$$

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

$$\begin{aligned} E(d_{it}d_{it-1}d_{it-2}\Delta\varepsilon_{it}|d_{is}, y^*_{is}) &= E(d_{it}d_{it-1}d_{it-2}(\Delta\varepsilon^0_{it} + \vartheta_0\Delta u_{it})|d_{is}, y^*_{is}) \\ &= E(d_{it}d_{it-1}d_{it-2}\vartheta_0(u_{it} - u_{it-1})|d_{is}, y^*_{is}). \end{aligned} \tag{18}$$

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}, conditional on d_{is} and y^*_{is}. Now, conditioning additionally on η_i and d_{it-2},

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

Notice that $d_{it}d_{it-1}(u_{it} - u_{it-1})$ is independent of d_{it-2}, d_{is}, and y^*_{is} conditional 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

$$\begin{aligned} E\left(d_{it}d_{it-1}(u_{it} - u_{it-1})|\eta_i\right) &= E\left(d_{it}d_{it-1}u_{it}|\eta_i\right) - E\left(d_{it}d_{it-1}u_{it-1}|\eta_i\right) \\ &= E\left(d_{it}u_{it}|\eta_i\right)E\left(d_{it-1}|\eta_i\right) - E\left(d_{it}|\eta_i\right)E\left(d_{it-1}u_{it-1}|\eta_i\right) = 0, \end{aligned} \tag{20}$$

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

$$E(s_{ist}y^*_{is}\Delta\varepsilon_{it}(\rho)) = (\rho_0 - \rho)E\left(s_{ist}y^*_{is}\Delta y^*_{it-1}\right). \tag{21}$$

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.

In sharp contrast with the case of the AB estimator, the system GMM estimator is not consistent. To illustrate this, we consider the unfeasible level moment conditions $E((y^*_{it}-\rho_0 y^*_{it-1})\Delta y^*_{it-1}) = 0$. The feasible analogue is $E(d_{it}d_{it-1}d_{it-2}(y^*_{it} - \rho_0 y^*_{it-1})\Delta y^*_{it-1})$ and we cannot guarantee that the expected value conditional on $d_{it} = d_{it-1} = d_{it-2} = 1$ equals 0. This condition implies that in the first stage equations $\eta_i + \gamma_0 z_{it} + u_{it} > 0$, $\eta_i + \gamma_0 z_{it-1} + u_{it-1} > 0$ and $\eta_i + \gamma_0 z_{it-2} + u_{it-2} > 0$. This implies that η_i and $u_{it}, u_{it-1}, u_{it-2}$, and, therefore, α_i and $\varepsilon_{it}, \varepsilon_{it-1}, \varepsilon_{it-2}$ are correlated.

Since d_{it} are discrete 0-1 variables, the events $\{d_{it} = 1, d_{it-1} = 1, d_{it-2} = 1\}$ and $\{d_{it}d_{it-1}d_{it-2} = 1\}$ are equivalent, and we have:

$0 = E[(y^*_{it} - \rho_0 y^*_{it-1})\Delta y^*_{it-1}] = E[E[(y^*_{it} - \rho_0 y^*_{it-1})\Delta y^*_{it-1} \mid d_{it}d_{it-1}d_{it-2}] = E[(y^*_{it} - \rho_0 y^*_{it-1})\Delta y^*_{it-1} \mid d_{it}d_{it-1}d_{it-2} = 1]P\{d_{it}d_{it-1}d_{it-2} = 1\} + E[(y^*_{it}-\rho_0 y^*_{it-1})\Delta y^*_{it-1} \mid d_{it}d_{it-1}d_{it-2} = 0]P\{d_{it}d_{it-1}d_{it-2} = 0\}$.

So, the expectation takes value 0 through a weighted combination of $E[(y^*_{it} - \rho_0 y^*_{it-1})\Delta y^*_{it-1} \mid d_{it}d_{it-1}d_{it-2} = 1]$ and $E[(y^*_{it} - \rho_0 y^*_{it-1})\Delta y^*_{it-1} \mid d_{it}d_{it-1}d_{it-2} = 0]$,

with probabilities $P\{d_{it}d_{it-1}d_{it-2} = 1\}$ and $P\{d_{it}d_{it-1}d_{it-2} = 0\}$ as weights.

Although the weighted combination is 0, we cannot ensure that any of its components is 0, so we cannot provide a bound for the bias of the estimator.

Our Monte Carlo experiments show that this is generally not equal to zero. However, these simulation exercises also show that $E(d_{it}d_{it-1}d_{it-2}(y^*_{it} - \rho_0 y^*_{it-1})\Delta y^*_{it-1})$ is, for all reasonable combination of the parameters of the model, very small and so is the induced bias (see Table C1 for an illustration).

The previous results for the AB estimator in the pure autorregressive model provide validity to the orthogonality restrictions of the first differenced equations, $E(\Delta\varepsilon_{it}y_{it-s}/z_i, d_{it} = d_{it-1} = d_{it-2} = 1) = 0; for \quad s \geq 2$. If we test the orthogonality restrictions of the level equations $E((\alpha_i + \varepsilon_{it})\Delta y_{it-1}/z_{it}, d_{it} = d_{it-1} = d_{it-2} = 1) = 0$, we have a standard Hansen/Sargan (see Sargan, 1988)

to check for sample selection.

3.2 Consistency in the dynamic model with covariates when $\delta = 0$

3.2.1 An exogenous covariate

We extend the previous AR(1) model to a model with a single exogenous covariate not included in the selection equation. The result can be straightforwardly generalised to many covariates.

$$y^*_{it} = \alpha_i + \rho_0 y^*_{it-1} + \beta_0' x^*_{it} + \varepsilon_{it}. \tag{22}$$

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. As before, we 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

$$E(s_{ist} y^*_{is} \Delta\varepsilon_{it}(\rho, \beta)) = (\rho_0 - \rho) E(s_{ist} y^*_{is} \Delta y^*_{it-1}) + (\beta_0 - \beta)' E(s_{ist} y^*_{is} \Delta x^*_{it}) + E(s_{ist} y^*_{is} \Delta\varepsilon_{it}), \tag{23}$$

and

$$E(s_{ivt} x^*_{iv} \Delta\varepsilon_{it}(\rho, \beta)) = (\rho_0 - \rho) E(s_{ivt} x^*_{iv} \Delta y^*_{it-1}) + (\beta_0 - \beta)' E(s_{ivt} x^*_{iv} \Delta x^*_{it}) + E(s_{ivt} x^*_{is} \Delta\varepsilon_{it}). \tag{24}$$

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

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

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,

$$
\begin{aligned}
E(s_{ivt}x^*_{iv}\Delta\varepsilon_{it}) &= E(d_{it}d_{it-1}d_{it-2}d_{iv}x^*_{iv}(\Delta\varepsilon^0_{it} + \vartheta_0\Delta u_{it})) \\
&= E(d_{it}d_{it-1}d_{it-2}d_{iv}x^*_{iv}\vartheta_0\Delta u_{it}) \\
&= E(d_{it-2}d_{iv}x^*_{iv}\vartheta_0 E(d_{it}d_{it-1}\Delta u_{it}|\eta_i, d_{it-2}, d_{iv}, x^*_{iv})) \\
&= E(d_{it-2}d_{is}x^*_{iv}\vartheta_0 E(d_{it}d_{it-1}\Delta u_{it}|\eta_i)) \\
&= 0. \qquad (25)
\end{aligned}
$$

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.

3.2.2 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 $v \leq t-2$. 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.

3.2.3 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

$$
E(s_{ist}y^*_{is}\Delta\varepsilon_{it}(\rho,\beta)) = (\rho_0 - \rho)E(s_{ist}y^*_{is}\Delta y^*_{it-1}) + (\beta_0 - \beta)'E(s_{ist}y^*_{is}\Delta x^*_{it}) + E(s_{ist}y^*_{is}\Delta\varepsilon_{it}), \quad (26)
$$

and

$$
E(s_{it}\xi_i\Delta\varepsilon_{it}(\rho,\beta)) = (\rho_0 - \rho)E(s_{it}\xi_i\Delta y^*_{it-1}) + (\beta_0 - \beta)'E(s_{it}\xi_i\Delta x^*_{it}) + E(s_{it}\xi_i\Delta\varepsilon_{it}). \quad (27)
$$

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 $E(s_{it}\xi_i\Delta\varepsilon_{it}) = 0$.

3.3 Consistency in the static model

All the aforementioned results hold when $\rho = 0$ and $x \perp z$. In particular when x is exogenous, there is no need to use an IV strategy and either the FE, FD or RE (GLS) estimators are consistent provided $cov(\alpha_i, x_{it}) = 0$. The proofs for the FE and FD estimators are straightforward, but we need to justify it for the RE estimator. Estimation of the uncorrected RE is carried out in the following selected sample:

$$y^*_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it} \qquad \text{if} \qquad d_{it} = 1, \tag{28}$$

where, under endogenous selection, $E(\alpha_i + \varepsilon_{it}|d_{it} = 1) = 0$, provided that $x \perp z$, x is independent of any transformation of z, in particular $\lambda(z)$. So, omission of the sample selection correction term does not affect the consistency of the estimate of β, a result which also applies to cross-sectional analysis.

In this static model we consider the extension in which x is not present in the selection equation but $x \not\perp z$. The uncorrected estimators are still consistent provided we control for the relationship between x and, say, Ψ, the covariates in z related to x. So, let us consider the following control function approach similar to Olsen's (1980) solution for sample selection in static models.

- Consider a vector of covariates $\Psi \in z$ such that $cov(x, \Psi) \neq 0$. Then, under standard assumptions, adding $E(x|\Psi)$ [or more generally $E(x|z)$] to the outcome equation corrects the bias. So, for the case of the static model estimated in levels, we adjust equation (28) as follows:

$$y_{it} = \alpha_i + \beta' x_{it} + \phi E(x_{it}|z_{it}) + m_{it} \quad \text{if} \quad d_{it} = 1,$$

where $m_{it} = \varepsilon_{it} + \phi E(x_{it}|z_{it})$.

- A simple test of the coefficient of $E(x|z)$, ϕ, evaluates the necessity of the correction.

This result can be applied to all models in which the covariates in both equations are distinct but not independent.

3.4 Consistency in models with covariates and $\delta \neq 0$

When at least one covariate is included in both the outcome and the selection equations, the uncorrected estimator is biased in the presence of endogenous sample selection. As suggested by Wooldridge (1995), bias correction induced by endogenous sample selection implies adding univariate selection terms if the sample is conditional on only one observation.

Result: Under the set of assumptions **B1** to **B3** for the first differenced equations and **B1** to **B3'** for the level equations (see Appendix B), Wooldridge's strategy can be extended to samples conditional on two observations (first-differenced models) and even to samples conditional on three consecutive observations (dynamic models) if the correlation structure is stationary and the time-variant errors are only contemporaneously correlated.

Alternatively, when these conditions fail to hold (also shown in Appendix B), we have to add bivariate corrections obtained from a bivariate probit model (first-differenced in static models and level equations in dynamic models) or from a trivariate probit model (first-differenced in dynamic models).

3.4.1 The correction procedure

We summarize the correction procedures in two steps (see Appendix B for details):

- **Step 1. Estimation of the selection equation**

  (i) **Errors contemporaneously correlated only under stationary correlation.** Under the assumption of normality of the errors in the selection equation, we estimate year-by-year probit models following the Mundlak/Chamberlain/Wooldridge approach and compute univariate correction terms. When x is fully exogenous, the specification includes the covariates z and x. We can solve any problem of correlation between these

covariates and the heterogeneity component by following either Mundlak or Chamberlain strategy, basically adding a correction for such correlation namely $g(z_i, x_i)$. Alternatively, when x is endogenous we replace x with current and lagged values of z.

(ii) **Serially cross-correlated errors.** We estimate bivariate probit models to correct equations in levels and first-differences for static models, or trivariate probit models to correct dynamic models. The order of the appropriate correction needed increases accordingly in AR(p) models (see Appendix B for details).

Important result: A follow up from cases where there is no need to correct is the fact that omission of any regressor in the selection equation, $\Psi \in z$, such that $\Psi \perp x$, does not affect the consistency of the corrected estimates.

- **Step 2. Estimation of the outcome equation**

(i) **Errors are only contemporaneously correlated.** In this case, assuming normality, all the estimators considered in this paper (FE, FD, RE, for the static model, and AH, AB, system GMM for the dynamic model) require corrections derived after adjusting univariate year-by-year probits. In the RE strategy and level equations of the system GMM estimator the corrections are introduced in levels. In first-differenced models, the corrections are introduced in first-differences. Finally, for the FE estimator, the correction is introduced using the within-transformation. For example, under the assumption that $x_{it} \perp \eta_i$, for level and first-differences equations in the dynamic case we have (see Appendix B for details and notation):

$$y_{it} = \rho y_{it-1} + x_{it}\beta + \sigma\lambda\left(H_{it}\right) + e_{it}. \tag{29}$$

$$\Delta y_{it} = \rho \Delta y_{it-1} + \Delta x_{it} + \sigma(\lambda\left(H_{it}\right) - \lambda\left(H_{it-1}\right)) + \Delta e_{it}, \tag{30}$$

where $H_{it} = z_{it}\gamma + \delta x_{it} + \bar{z}_i\theta$ and $e_{it} = \varepsilon_{it} + \lambda\left(H_{it}\right)$.

(ii) **Serially cross-correlated errors under stationary correlation.** As described in Appendix B, in static models estimated by GLS (RE) we only need to add a single correction; in static models estimated by FD we need to add two correction terms ob-

tained from a bivariate probit (evaluating the expectation of the first-differenced error conditional on two errors of the selection equation). In dynamic models estimated using the AH or the AB estimator, we need to add at least two correction terms obtained from a trivariate probit (evaluating the expectation of the first-differenced error conditional on the errors of the selection equation in the current, lagged and lagged twice periods). Finally, when obtaining the system GMM estimator we combine the solution for the AB estimator (trivariate corrections) with the solution offered for the level model estimated in first differences. This means that the correction to the level and first differenced equations is not the same, so the estimator cannot be obtained using standard software (as xtabond2 in Stata, for instance).

We provide the corrections needed for the system GMM estimation as an example. We note that when x is endogenous the corrections need to be instrumented using the same lag order used to instrument the covariate (details are provided in Appendix B).

$$y_{it} = \rho y_{it-1} + x_{it}\beta + \bar{w}_i\psi + \sigma_0\lambda\left(H_{it}, H_{it-1}, \varrho_{t,t-1}\right) + \sigma_{-1}\lambda\left(H_{it-1}, H_{it}, \varrho_{t,t-1}\right) + e_{it}. \quad (31)$$

$$\begin{aligned}\Delta y_{it} = {} & \rho\Delta y_{it-1} + \Delta x_{it} + \bar{\sigma}(\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) \\ & -\lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right)) \\ & +\bar{\sigma}_{-2}\lambda\left(H_{it-2}, H_{it-1}, H_{it}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) + \Delta e_{it},\end{aligned} \quad (32)$$

where $\varrho_{t,t-s}$ denotes the correlation between errors in period t and $t-s$ and the functions involving H and ϱ (the selection corrections) are defined in Appendix B.

In all cases, it is necessary to compute corrected standard errors. This can be done by means of the delta method or bootstrapping. Finally, one can use a standard t-test for the significance of the correction term, or a Wald test in case of multiple lambda's (Wooldridge, 1995).

3.4.2 Construction of the corrections

For a typical static selection model, as described in equation (2), and assuming, for simplicity, normality of $\eta_i + u_{it} = \nu_{it}$, we estimate a probit for each period and then compute the well-known

selection term $\hat{\lambda}_{it}(w_{it}\hat{\gamma})$. When we allow correlation between w_{it} (w stands for the combination of z and x) and η_i, we can rely on Mundlak (1978) and assume, for instance, $\eta_i = \tilde{w}_i\varphi$, where $\tilde{w}_i$ is the vector of individual means of w_{it}, and we, again, can estimate a probit for each period and compute $\tilde{\lambda}_{it}(z_{it}\tilde{\gamma} + \tilde{w}_i\tilde{\varphi})$, which is then introduced in the second step as before.

In the case of a dynamic selection equation, the lagged observed regressor is correlated with the random effect by construction. If 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 $\eta_i = \pi_1 w_{i1} + \pi_2 w_{i2} + ... + \pi_T w_{iT}$ and recover the corresponding selection terms. However, strictly speaking, to recover the structural parameters of the selection equation, we should estimate a probit model for each year based on a reduced form, where d^*_{it} is a function of all exogenous variables (i.e., z) and we predict the index $\hat{d}^*_{it}$. Then, in a second stage, we estimate the structural parameters by within-groups, Minimum Distance or GMM and compute the correction terms based on these two-stage coefficients (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.

The previous univariate corrections do not work if the errors $\epsilon_{it}, \nu_{it}, \nu_{it-1}, \nu_{it-2}$ are jointly normal. In this case, we can estimate bivariate or trivariate probits in order to construct the bivariate and trivariate corrections. In Appendix B we provide additional details as well as semiparametric estimates of the correction that can overcome the failure of the normality assumption (see also Rochina-Barrachina, 1999, Gayle and Viauroux, 2007 and Jiménez-Martín *et al.*, 2009).

4 Monte Carlo experiments

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

$$d^*_{it} = a - z_{it} - \delta x_{it} - \eta_i - u_{it}, \tag{33}$$

and

$$d_{it} = 1[d^*_{it} > 0], \tag{34}$$

where a is set so that $p(d_{it}^* > 0) = 0.85$ and $z_{it} \sim N(0, \sigma_z)$ with $\sigma_z = 1$. Note that when $\delta = 0$, x is not present in the selection equation. Second, the outcome of interest is generated as follows:

$$y_{it}^* = (2 + \beta x_{it} + \alpha_i + \varepsilon_{it})/(1 - \rho) \text{ if } t = 1, \tag{35}$$

$$y_{it}^* = 2 + \rho y_{it-1}^* + \beta x_{it} + \alpha_i + \varepsilon_{it} \text{ if } t = 2, ..., T, \tag{36}$$

and

$$y_{it} = y_{it}^* \text{ if } d_{it} = 1. \tag{37}$$

We let ρ vary between 0 (static model), 0.25, 0.50 and 0.75. We generate all variables for $T = 1$ to $T = 20$ and discard the first 13 observations to minimize the effects of the initial conditions. The results remain unchanged if we use these extra 13 observations and, thus, start the observed sample with an initial condition for each individual in the sample. We consider the following process for x:

$$x_{it} = (0.5 + \wp_{it} + \alpha_i^x + \varepsilon_{it}^x + \kappa_1 \alpha_i + \kappa_2 \varepsilon_{it}))/0.5 \quad \text{if } t = 1, \tag{38}$$

and

$$x_{it} = 0.5 + 0.5 x_{it-1} + \wp_{it} + \alpha_i^x + \varepsilon_{it}^x + \kappa_1 \alpha_i + \kappa_2 \varepsilon_{it}) \quad \text{if } t > 1. \tag{39}$$

and we let κ_2 vary between $\kappa_2 = 0$, that is x is fully exogenous, and $\kappa = 0.5$, which implies x is either endogenous or predetermined (in which case ε_{it} is replaced by ε_{it-1}). For ease of exposition, we assumed that $\kappa_1 = 0$ except when estimating the static level equation by GLS, where we also consider the case $\kappa_1 = 0.5$. We further assume that $cov(x_i, \eta_i) = 0$ and $cov(z_i, \eta_i) = 0$. Removing these assumption (that only affect level-based estimates) does not affect any of the relevant results in the paper. Simulations removing these assumptions are available for the static model case.

Finally, we assume the following structure for z, $\wp$ as well as the errors:

$$\wp_{it} \sim N(0, \sigma_{\wp}) \text{ with } \sigma_{\wp} = 1, \tag{40}$$

$$z_{it} \sim N(0, \sigma_z) \text{ with } \sigma_z = 1, \tag{41}$$

$$\eta_i \sim N(0, \sigma_\eta) \text{ with } \sigma_\eta = 1, \tag{42}$$

$$u_{it} \sim N(0, \sigma_u) \text{ with } \sigma_u = 1, \tag{43}$$

$$\alpha_i = \alpha_i^0 + 0.5\eta_i, \alpha_i^0 \sim N(0, \sigma_{\alpha^0}) \text{ with } \sigma_{\alpha^0} = 1, \tag{44}$$

$$\varepsilon_{it} = \varepsilon_{it}^0 + \vartheta_0 u_{it} + \vartheta_1 u_{it-1} + \vartheta_2 u_{it-2}, \varepsilon_{it}^0 \sim N(0, \sigma_{\varepsilon^0}) \text{ with } \sigma_{\varepsilon^0} = 1, \tag{45}$$

$$\alpha_i^x \sim N(0, \sigma_{\alpha^x}) \text{ with } \sigma_{\alpha^x} = 1, \tag{46}$$

and

$$\varepsilon_{it}^x \sim N(0, \sigma_{\varepsilon^x}) \text{ with } \sigma_{\varepsilon^x} = 1, \tag{47}$$

where, in the case A1 of contemporaneous correlation, we set $\vartheta_0 = 0.5; \vartheta_1 = \vartheta_2 = 0$. These assumptions imply that $corr(\varepsilon_{it}, u_{it}) = corr(\alpha_i, \eta_i) = 0.5/\sqrt{1 + 0.5^2} = 0.447$. Alternatively, in the case of serially cross-correlated errors we set $\vartheta = 0.5; \vartheta_1 = 0.5/2; \vartheta_2 = -0.5/3$.

4.1 Description of the experiments

For each experiment, we set the initial (before selection) 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. In dynamic models we need at least three consecutive observations of the same regime 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 degree of sample selection. For example, a 15 per cent of initial selection implies loosing around 1/3 of the observations. In static models with exogenous regressors this loss is not important. For each combination of the parameters we perform 500 replications.

Under the assumption of contemporaneous correlated errors, we simulate the following five combinations of the parameters of interest, linked to the cases already described in Table **1**:

(i) Static model with an exogenous x (with and without correlation with z) not present in the selection equation: $\rho = 0$, $\beta = 1$ $\delta = 0$

(ii) Static model with an exogenous x also present in the selection equation: $\rho = 0$, $\beta = \delta = 1$

(iii) Purely AR(1) model: $\rho = 0.25, 0.50, 0.75$, $\beta = \delta = 0$

(iv) Dynamic model with an endogenous covariate either present or not in the selection equation and contemporaneously correlated time varying errors : $\rho = 0.25; \rho = 0.75$, $\beta = 1$, $\delta = 0$ or $\delta = 1$, $(cov(\varepsilon_{it}, u_{is}) \neq 0; s < t)$

(v) Dynamic model with an exogenous covariate either present or not in the selection equation and serially cross-correlated time varying errors : $\rho = 0.25; \rho = 0.75$, $\beta = 1$, $\delta = 0$ or $\delta = 1$, $cov(\varepsilon_{it}, u_{is} \neq 0; s \leq t)$

In each case, we evaluate the performance of the appropriate estimators as described in Table **1**. In (i) and (ii) we evaluate the FE, FD and RE estimators. In (iii) to (v) we evaluate two GMM estimators: AB and system GMM. Selection of the instruments is done 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 as well as current values and lags of the exogenous regressors. Although we are aware of the instrument proliferation issue analyzed by Roodman (2009), it does not constitute a problem here given the reduced number of periods (a maximum of 7) remaining for estimation, but we also use Roodman's proposal to collapse the number of instruments and we get very similar results.

4.2 Simulation results

Although we have simulated the five combinations previously considered, we are going to present in this section only the most interesting results and we relegate the rest of results to Appendix C for interested readers.

Simulations of static models (with an exogenous regressor) either with $\delta = 0$ and $x \perp z$, $\delta \neq 0$ and $x \perp z$ or $\delta = 0$ and $x \not\perp z$, all of them under the assumption that the errors in both equations are contemporaneously correlated are given in Table C2 in Appendix C. The results for $\delta = 0$ and $x \perp z$, that is, in the case that correction is not needed, show that the average bias is almost zero, regardless of the sample size. According to the RMSE criterion (and also confidence interval coverage rates

(CICR)), since $cov(x, \alpha_i) = 0$, the RE (either with or without correlated covariates) is our preferred method, as expected. In the scenario $\delta \neq 0$ and $x \perp z$ we show again that the RE estimator is the preferred option, attending the RMSE criterion (CICR as well), provided $cov(x, \alpha_i) = 0$, but the uncorrected estimates are biased because of the presence of common variables in both equations. However, the bias is small and we observe minor differences when including correction terms *a la Wooldridge*. The most interesting case (reported in panel C) in the static model arise when $\delta = 0$ and $x \not\perp z$ and x is not included in the selection equation but is correlated with some variables included in the vector z, say Ψ. All the uncorrected estimates are biased, but instead of including correction terms *a la Heckman*, we get almost complete bias reduction if we add to the outcome equation an estimate of $E(x|\Psi)$, especially as the sample size grows.

In all simulations we report the empirical rejection frequency (ERF) of the sample selection test corresponding to the corrected estimator under the null hypothesis that the selection term is not necessary in the outcome equation. The ERF computes the percentage of rejection of the null in 500 replications. When there is endogenous selection (the null is false) and the initial N is small, we reject both the FE and the RE estimators in 95% and 99.8% of the cases, respectively, while the rejection rate of the FD estimator is smaller, 76.8%. When the sample is large (N is 5000) we always reject the null. When the null is true (no endogenous selection) we reject the null in between 3.8% (FE estimator and N large) and 7% (RE and N small) of the cases.

Another set of simulation results that deserves some explanation refer to the experiments with a pure AR(1) small-T (max $T = 7$) panel data model. Up to 20 observations are simulated for each case, the initial sample is obtained after discarding the first 13 observations for each individual. See Table C3 in the appendix for simulation results using the initial value for each individual and up to the next six observations of the process for each individual. We present in the main text, see Table **2**, simulations for different values of the autoregressive parameter under the assumption that the errors are only contemporaneously correlated, estimating the AB and the system GMM estimators under alternative 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 on average). In the case of results without endogenous selection when the initial sample is small ($N = 500$) the bias of the AB grows with the autoregressive parameter and becomes sizable from $\rho = 0.75$ (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). As we increase the sample size ($N = 5000$), the average bias of the AB estimator is reduced substantially and remains noticeable only for $\rho > 0.75$. The system GMM estimator, which is consistent in this case, shows a very small bias for $N = 500$ (never exceeding one per cent), and even smaller when $N = 5000$. Figure C1 in Appendix C confirms these results with a sample size varying from $N = 200$ to $N = 5000$ in the absence of any sort of selection (estimators labeled AB all and system all).

When endogenous sample selection is considered, we do not detect any significant change in the biases for the uncorrected AB estimator for both selection models. Even when the initial sample is small, the difference between the cases with and without selection is practically undetectable (although the smaller effective sample size in the selected sample leads to higher RMSE). In contrast, the system GMM estimator always shows a very small bias (between 1 per cent for $\rho = 0.25$ and 2.25 per cent for $\rho = 0.75$). In terms of RMSE and CICR, when the sample size is small ($N = 500$), they favor the system estimator. However, when the sample size grows ($N = 5000$) the choice under both criteria is reversed. In fact, with a much large sample size ($N = 50000$), the choice under both criteria is much clearer.

Table 2: Average bias, RMSE and CICR in the purely AR(1) model. T=7 (after discarding the first 13 generated observations); 500 replications

	Estimates with the full sample						Estimates with the selected sample					
	AB estimator			system			AB estimator			system		
ρ	av. bias	RMSE	CICR[1]	av. bias	RMSE	CICR	av. bias	RMSE	CICR	av. bias	RMSE	CICR
Panel A: N=500												
0.25	-0.0057	0.0407	0.940	0.0005	0.0320	0.940	-0.0141	0.7660	0.954	-0.0034	0.7546	0.938
0.50	-0.0117	0.0560	0.944	0.0020	0.0368	0.936	-0.0317	0.5372	0.950	-0.0082	0.5108	0.944
0.75	-0.0425	0.1014	0.944	0.0084	0.0445	0.922	-0.1003	0.3742	0.870	-0.0091	0.2670	0.950
Panel B: N=5000												
0.25	-0.0013	0.0119	0.952	-0.0003	0.0094	0.956	-0.0011	0.7513	0.946	-0.0037	0.7538	0.942
0.50	-0.0020	0.0163	0.946	-0.0001	0.0113	0.946	-0.0025	0.5031	0.940	-0.0095	0.5097	0.930
0.75	-0.0045	0.0290	0.938	0.0007	0.0137	0.946	-0.0086	0.2620	0.942	-0.0182	0.2689	0.880

1. CICR. 95 % Confidence intervals coverage rates, ie. CICR= $\sum_S \mathbf{1}(\hat{a} - 1.96 * s.e.(\hat{a}) < a < \hat{a} + 1.96 * s.e.(\hat{a}))/S$, where S is the number of simulations, a is the true parameter and $\hat{a}$ is the estimate of a in each simulation.

Some additional conclusions can be drawn when varying the sample size (Figure C1 in Appendix C). When $N = 200$, the AB estimator shows sizable bias, which decreases as N increases. The

system GMM estimator has always very small bias, however. For a given ρ, it remains stable (between 1 and 2.5 per cent) as N increases. We detect a threshold for N for each combination of parameters, the average bias of the system GMM estimator being smaller below this threshold, and larger above it. Therefore, we conclude that for moderate and small sample sizes (say, below the range 1000-1500), the system GMM estimator is highly recommended because of the likely smaller bias as well as smaller variance. Finally, when α_i and η_i are not correlated, the bias of the system GMM estimator tends to disappear (in comparison with the previous case) due to the fact that the main source of bias is the correlation between the heterogeneous components of the outcome and selection equations (Table C1 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 correlation between the error components and the autoregressive parameter).

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 GMM estimator (especially when the available sample size is small), one is tempted to use a simple procedure as the one described for the static model following Olsen (1980). However, we should emphasize that methods based on OLS in dynamic models can only be used as bias reduction approaches (see Han and Lee, 2022) because as it is well-known it does not provide consistent estimates in linear probability models as shown by Horrace and Oaxaca (2006).

In addition to the experiments above, we have carried out several Monte Carlo exercises with cases departing from the basic assumptions of the purely AR(1) model we have simulated (with the exception of the initial conditions case which is reported in Table C3 in Appendix C, these results are not reported in the paper, but they are available upon request from the authors). The following robust checks were performed: (a) In the first panel of Table C3 we present the same experiments reported in Table **2** using the first seven realizations of the process for each individual (that is without discarding the first 13 observations). We also used the Han and Phillips (2010) estimator, which does not suffer from weak/many instruments problem and works very well regardless of the magnitude of AR(1) coefficient to check the sensitivity of our results to alternative dynamic panel data estimators that perform well when the stationarity assumption is not satisfied; (b) varying the longitudinal dimension of the panel; (c) increasing the percentage of selection (from 0.15 to 0.25);

(d) increasing the ratio of the variances to $\frac{\sigma_\alpha^2}{\rho_\varepsilon^2} = 2$; (e) reducing the correlation between the errors (the correlation parameter is reduced from 0.5 to 0.25); (f) and, finally, introducing 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 by multiplying either ε_{it} or u_{it} by a time-varying Bernoulli process taking the values 1 or 2. We also allow the correlation coefficient between the time-varying errors in (1) and (2) to vary over time by multiplying ϑ by either 0.5, 1 or 2. All these sensitivity exercises confirm the main lessons drawn from the previous analysis: the AB (or the AH) estimator is moderately biased when N is small or moderate, and unbiased when N is large. These additional results by and large recommend the system GMM estimator for the small N case and the AB for the large N case.

We performed additional Monte Carlo exercises for dynamic models with a covariate that is either present or absent from the selection equation. This variable can be either exogenous, predetermined or endogenous. The key results obtained with an endogenous covariate and contemporaneously cross-correlated errors are shown in the first two columns (for $\rho = 0.25$ and $\rho = 0.75$) of Panels A and B in Table **3**. The small biases found for the AB estimator with $N = 500$ decrease as sample size increases (they practically disappear when $N = 5000$). Note that the CICR criteria also show that the AB is appropriate. The system GMM estimator, although not consistent, has a very small bias regardless of the sample size. More importantly, the RMSE is smaller (and the CICR is similar in magnitude) than for the AB, even when the sample is large ($N = 5000$). Note however, that for very large samples (say, $N = 50000$, results not reported) the latter remark is no longer true, since the bias of the AB estimator goes to zero while the bias of the system GMM estimator does not. All these results also apply to the case where x is predetermined or exogenous. We do not report them, but they are available upon request from the authors.

Next, we focus on a dynamic model with a covariate x present in both, the outcome and the selection equations (simulation results for this case when x is not present in the selection equation and is not correlated with z are available upon request). We present both uncorrected and corrected estimates and $cov(\varepsilon_{it}, u_{is} = 0); s < t$. The uncorrected results are reported in the third and fourth row and the corrected ones in the next two rows of Panels A and B in Table **3**. The uncorrected estimates are biased regardless of the sample size, which shows the need to correct for sample selection when there is at least a common covariate in both equations. Given

Table 3: Average bias, RMSE and CIRC in the dynamic model with an endogenous covariate. ($cov(\varepsilon_{it}, u_{is}) \neq 0; s < t$) T=7; 500 replications

			AB							SYSTEM						
x in selection	Corrected	value ρ	ρ av. bias	ρ RMSE	ρ CICR	β av. bias	β RMSE	β CICR	λ test ERF	ρ av. bias	ρ RMSE	ρ CICR	β av. bias	β RMSE	β CICR	λ test ERF
Panel A: N=500; endogenous selection																
No	No	.25	-0.0109	0.0323	.94	0.0098	0.0393	0.914		-0.0060	0.0239	.944	0.0074	0.0327	0.934	
No	No	.75	-0.0291	0.0591	.896	-0.0087	0.0505	0.956		-0.0051	0.0224	.952	0.0043	0.0305	0.926	
Yes	No	.25	-0.0236	0.0376	.868	-0.0389	0.0575	0.876		-0.0102	0.0282	.926	-0.0368	0.0551	0.848	
Yes	No	.75	-0.0334	0.0465	.824	-0.0551	0.0714	0.812		-0.0155	0.0357	.896	-0.0441	0.0610	0.824	
Yes	Yes[1]	.25	-0.0344	0.0451	.76	0.0235	0.0547	0.924	0.510	-0.0184	0.0321	.896	0.0129	0.0480	0.934	0.420
Yes	Yes[1]	.75	-0.0387	0.0505	.768	-0.0003	0.0529	0.944	0.454	-0.0152	0.0331	.908	0.0044	0.0459	0.938	0.430
Panel A: N=500; exogenous selection																
Yes	Yes[1]	.25	-0.0165	0.0383	.906	0.0007	0.0470	0.950	0.038	0.0003	0.0309	.938	0.0009	0.0388	0.966	0.040
Yes	Yes[1]	.75	-0.0245	0.0462	.894	-0.0132	0.0539	0.934	0.040	0.0135	0.0317	.91	0.0023	0.0392	0.960	0.052
Panel B: N=5000; endogenous selection																
No	No	.25	-0.0011	0.0101	.936	0.0007	0.0109	0.962		-0.0020	0.0078	.934	0.0007	0.0093	0.954	
No	No	.75	-0.0041	0.0165	.934	-0.0020	0.0148	0.964		-0.0047	0.0084	.908	-0.0014	0.0091	0.950	
Yes	No	.25	-0.0128	0.0158	.666	-0.0553	0.0568	0.010		-0.0053	0.0100	.898	-0.0493	0.0508	0.022	
Yes	No	.75	-0.0217	0.0239	.394	-0.0637	0.0651	0.004		-0.0179	0.0207	.564	-0.0562	0.0577	0.010	
Yes	Yes[1]	.25	-0.0192	0.0212	.4	-0.0076	0.0168	0.932	1.000	-0.0114	0.0140	.666	-0.0049	0.0144	0.942	1.000
Yes	Yes[1]	.75	-0.0233	0.0253	.326	-0.0191	0.0248	0.772	1.000	-0.0148	0.0177	.642	-0.0100	0.0171	0.882	1.000
Panel B: N=5000; exogenous selection																
Yes	Yes[1]	.25	-0.0020	0.0108	.948	-0.0005	0.0139	0.958	0.058	-0.0003	0.0088	.944	-0.0008	0.0124	0.950	0.050
Yes	Yes[1]	.75	-0.0028	0.0117	.954	-0.0023	0.0156	0.944	0.062	0.0018	0.0109	.95	-0.0004	0.0129	0.942	0.044

1. In Panels A and B the correction is obtained from a year by year probit with z and $\wp$ as covariates.

that $cov(\varepsilon_{it}, u_{is} = 0); s < t$, for GMM-IV estimators, as we show in the Appendix B, this implies adding univariate correction terms (*a la Wooldrigde*) to each equation (first differenced corrections in first differenced equations for both the AB and system estimators, and level corrections in level equations for the system estimator). Furthermore, since x is endogenous, these additional terms need to be instrumented using backward lags. The effects of sample correction on the magnitude of the bias reduction is sizable, especially for β. Reductions in the RMSE are related to the sample size. When the sample size is small the ERF is small (around 0.50), so the sample selection test fails to clearly detect the presence of endogenous sample selection for both estimators. As the sample size increases the performance of the test improves substantially with an ERF close to 1. When the null is true the ERF has a range of 0.38 (highest) to 0.06 (lowest).

The simulation exercise reported in Table **4** explores a model with a single exogenous covariate (x and z, respectively) in each equation and time-variant errors that are cross-serially correlated $cov(\varepsilon_{it}, u_{is} \neq 0); s \leq t$. We want to stress the fact that when the two equations do not have common covariates and they are independent, there is no need to correct the estimates, even when the correlation structure is very complex. The results from this experiment are reported in the first two rows of panels A and B in Table **4**. In the third and fourth row of panels A and B we report the case in which x is present in both equation and we do not correct for sample selection. The

simulation results in Table **4** are in line with prior expectations since the bias of the uncorrected estimator is sizable, especially for β, a feature shared by many of the results we have presented so far, and it does not decrease as N grows.

Finally, in the remaining rows of panels A and B, we report corrected estimates. First think to note is the fact in models were the time-variant errors are cross-serially correlated, i.e., $cov(\varepsilon_{it}, u_{is} \neq 0); s \leq t$, and they have an exogenous covariate present in both equations, we show in Appendix B that the estimation of the model either by FD-GMM or system GMM requires multiple correction terms. As shown in appendix B, we have to add two correction terms obtained from trivariate probit models for the first-differenced equations (present in both the AB and the system estimators) and two additional terms obtained using bivariate probit models for the equation in levels. Moreover, given the multiplicity of correction terms, we have to use a Wald test instead of a typical t-test to check for sample selectivity.

When the null of endogenous selection is true (reported in rows fifth and sixth of panels A and B) the bias of the corrected estimator is very small and decreases with N. Likewise, the CICR statistic is found around 0.95 in a majority of cases, being the case of CIRC statistic for ρ in the corrected system estimation a notable exception. On the other hand, the ERF of the correction terms is moderate when N is small and increases to a value close to 1 as N grows. Alternatively, when the null of endogenous selection is not true the ERF of the sample selection test stabilizes between 0.06 ($N = 500$) and 0.04 ($N = 5000$) both for the AB and system GMM estimators.

Our final Monte Carlo exercise compares univariate tests of selection bias presented in Panels A and B of Table **4** with multivariate ones. In the presence of sample selection but absence of longitudinal cross-correlation between the outcome and the selection, i.e., $cov(\varepsilon_{it}, u_{it} \neq 0)$ and $cov(\varepsilon_{it}, u_{is} = 0; s < t)$, we simulate the GMM estimators with two correction terms. Wooldridge-like corrections are adequate (Heckman's lamdba in first differences and levels in the first-differenced and in the levels equations, respectively). In these cases, it is easy to show that the coefficient of the lagged twice trivariate correction term in the first-differenced equations and the coefficient of the lagged bivariate lambda in the equation in levels should be zero. Then, a simple t-test for the corrected AB estimator or a Wald test for the corrected system GMM estimator stand as checks for the longitudinal correlation between the errors in the outcome and the selection equations. We obtain the expected results as reported in Panel C of Table **4**.

Table 4: Average bias, RMSE and CICR in the dynamic model with an exogenous covariate. $cov(\varepsilon_{it}, u_{is} \neq 0; s \leq t)$ T=7; 500 replications

			AB							SYSTEM						
x in selection	Corrected	value ρ	av. bias	ρ RMSE	CICR	av. bias	β RMSE	CICR	λ's test ERF	av. bias	ρ RMSE	CICR	av. bias	β RMSE	CICR	λ's test ERF
Panel A: N=500; endogenous selection																
No	No	.25	-0.0004	0.0225	0.952	-0.0032	0.0268	0.940		0.0174	0.0274	0.876	0.0068	0.0269	0.924	
No	No	.75	-0.0132	0.0257	0.902	-0.0051	0.0277	0.926		0.0078	0.0198	0.910	0.0108	0.0276	0.918	
Yes	No	.25	-0.0193	0.0345	0.900	-0.0402	0.0506	0.750		0.0023	0.0277	0.946	-0.0316	0.0442	0.824	
Yes	No	.75	-0.0266	0.0377	0.824	-0.0447	0.0545	0.714		-0.0089	0.0257	0.912	-0.0354	0.0475	0.800	
Yes	Yes	.25	-0.0047	0.0412	0.940	-0.0020	0.0450	0.938	0.322	0.0046	0.0380	0.946	-0.0035	0.0436	0.962	0.170
Yes	Yes	.75	-0.0217	0.0474	0.880	-0.0109	0.0477	0.946	0.312	-0.0175	0.0315	0.898	-0.0094	0.0442	0.966	0.154
Panel A: N=500; exogenous selection																
Yes	Yes	.25	-0.0112	0.0387	0.922	-0.0012	0.0398	0.952	0.062	0.0002	0.0335	0.940	-0.0003	0.0378	0.962	0.070
Yes	Yes	.25	-0.0124	0.0384	0.920	-0.0043	0.0421	0.948	0.070	0.0052	0.0225	0.932	0.0018	0.0390	0.958	0.076
Panel B: N=5000; endogenous selection																
No	No	.25	0.0061	0.0094	0.852	-0.0015	0.0075	0.960		0.0169	0.0182	0.264	0.0065	0.0096	0.876	
No	No	.75	-0.0050	0.0086	0.864	-0.0019	0.0077	0.950		0.0039	0.0073	0.868	0.0095	0.0118	0.754	
Yes	No	.25	-0.0098	0.0129	0.784	-0.0386	0.0395	0.016		0.0024	0.0084	0.940	-0.0325	0.0336	0.044	
Yes	No	.75	-0.0151	0.0171	0.504	-0.0401	0.0411	0.012		-0.0158	0.0173	0.374	-0.0386	0.0396	0.018	
Yes	Yes	.25	0.0074	0.0137	0.918	-0.0011	0.0134	0.966	0.998	0.0004	0.0106	0.940	-0.0076	0.0153	0.912	0.962
Yes	Yes	.75	-0.0043	0.0130	0.924	-0.0044	0.0144	0.940	0.998	-0.0259	0.0269	0.068	-0.0131	0.0185	0.826	0.954
Panel B: N=5000; exogenous selection																
Yes	Yes	.252	-0.0020	0.0104	0.956	0.0003	0.0113	0.958	0.048	-0.0010	0.0092	0.966	0.0003	0.0113	0.942	0.044
Yes	Yes	.752	-0.0021	0.0103	0.954	-0.0003	0.0118	0.956	0.042	0.0002	0.0062	0.956	0.0003	0.0115	0.950	0.040

Testing univariate corrections vs multiple corrections

			AB							SYSTEM						
x in selection	Corrected	value ρ	av. bias	ρ RMSE	CICR	av. bias	β RMSE	CICR	$xtra\lambda$'s test ERF	av. bias	ρ RMSE	CICR	av. bias	β RMSE	CICR	$xtra\lambda$'s test ERF
Panel C1: N=500; endogenous selection but $cov(\varepsilon_{it}, u_{is} = 0; s < t)$																
Yes	Yes	.25	-0.0147	0.0433	0.916	-0.0017	0.0438	0.946	0.306	-0.0025	0.0374	0.944	-0.0010	0.0418	0.952	0.090
Yes	Yes	.75	-0.0195	0.0452	0.896	-0.0082	0.0468	0.934	0.318	0.0018	0.0253	0.946	-0.0014	0.0422	0.956	0.086
Panel C2: N=5000; endogenous selection but $cov(\varepsilon_{it}, u_{is} = 0; s < t)$																
Yes	Yes	.25	-0.0030	0.0115	0.954	-0.0007	0.0132	0.954	0.990	-0.0054	0.0113	0.920	-0.0022	0.0132	0.952	0.048
Yes	Yes	.75	-0.0036	0.0122	0.930	-0.0015	0.0136	0.948	0.988	-0.0053	0.0090	0.888	-0.0032	0.0134	0.948	0.038

1: In Panels A to C the correction is obtained from trivariate probits (for FD equations) and bivariate probits (for level equations) with z, $z(-1)$ and $z(-2)$ as covariates (in the trivariate case) or z, $z(-1)$ in the bivariate one.

5 Empirical applications

This section presents two applications of the proposed methods. The first uses well-known data from the Panel Study of Income Dynamics (PSID) to estimate log hourly earnings equations of US females. This dataset has been employed in several empirical papers with different purposes, but we use it to compare our results to alternative methods for selection models proposed by Semykina and Wooldridge (SW). The second uses consumption data from the Spanish Continuous Family Expenditure Survey (ECPF from now on) to adjust myopic models of tobacco consumption. This is the same dataset used by Jones and Labeaga (2003). They were worried about the censoring nature of the observations and how to handle it in the framework of a rational addiction model of tobacco consumption (see Becker and Murphy, 1988). Our objective here is to estimate a myopic model of consumption trying to mimic our autoregressive proposal.

5.1 Estimating female earnings equations

In this first application, we employ the same data used in SW, which were also used by Lai and Tsai (2018) (we compare our results with those presented by SW, but, unfortunately, we cannot compare with Lai and Tsay, 2018, because they estimated a static sample selection model). The data consists of a panel taken from the PSID covering the period 1980-1992, and we use the same selection rules (see Section 6 in SW). Since we discuss pure autoregressive models in the paper, we estimate it on this data and we present the results in Table C4 of Appendix C. We extend the model in Table **5** to include age, age squared and number of years of education. These variables together with family size are included in the selection equation. In terms of the notation used in (3), age, age squared and number of years of education form the vector x_{it} and family size, which is an exclusion restriction, is included in z_{it}. In the case of family size, we include, as SW, z_{i1}, z_{i2}, ..., z_{iT}. The first column in Table C4 presents first-differenced IV estimates. Alternatively, column (1) in Table **5** reports the SW estimator. Columns (2) and (3) in both tables report the AB and system GMM results obtained in the selected sample when we do not correct the earnings equation. Alternatively, in columns (4) and (5) of Table **5** and column (4) of Table C4 we present year-by-year probit corrections under the assumption that the errors in both equations are contemporaneously correlated.

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. 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 finds that the AB and system GMM results are close. 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.

In Table **5**, 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 equations in levels and first-differences. The number of overidentifying restriction is 65 in the 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 (Roodman, 2009). When we use up to the fourth lags 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 **5**. The autoregressive coefficient (as well as its standard error) remains very similar in the extended model in Table **5** compared to the pure autoregressive case, and it is substantially lower than the one obtained by SW. Given that all first stage variables are either time-invariant (education) or deterministic (age and age square) the uncorrected first differences estimates are consistent.

The proposed corrections of the system GMM estimator do not imply significant changes in the key coefficients of the model. All in all, 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 Arellano *et al.*, 1999).

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 is more in line with the average return to education for females for the US usually found in the literature (see Card, 1999 or Harmon *et al.*, 2003). We do not detect endogenous selection due to the correlation between the time-invariant heterogeneity components (column (5) and (6) in Table **5**). The coefficient of lag of log hourly earnings in SW and in our application are different. Our guess is that the specification estimated by SW does not control adequately for the correlation between the fixed effects and the lagged dependent variable (remember that they estimate the model by pooled NLS or GMM). Our estimator controls for the fixed effects by first differencing. Moreover, the addition of the equation in levels helps in identifying the effects of education, age and age squared and improves the efficiency of the estimates of these coefficients.

All in all, our opinion is that the similarities among the coefficients with and without correcting for selectivity are in line with 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

Table 5: **Estimates for the dynamic log hourly earnings equation with covariates**

	(1)	*(2)*	*(3)*	*(4)*	*(5)*
	Semykina Wooldridge	*No correction*	*No correction*	*year by year correction first dif eq*	*year by year correction all equations*
	GMM	*AB*	*system*	*AB*	*system*
Lag log	0.5740***	0.1047**	0.1850***	0.1170***	0.2189**
hourly earnings	(0.0400)	(0.0374)	(0.0436)	(0.0379)	(0.0447)
Education	0.0290***	—	0.0949***	—	0.0931***
	(0.004)		(0.0084)		(0.0085)
Age	0.0090***	0.0070	0.0375***	0.0269**	0.0228***
	(0.004)	(0.0127)	(0.0113)	(0.0128)	(0.0126)
Age squared	-0.0001***	-0.0001	-0.0004***	-0.0001	-0.0003***
	(0.000)	(0.0001)	(0.0001)	(0.0002)	(0.0001)
Observations	5033	5033	5033	5033	5033
Joint significance	41.3 (10)	–	–	11.27 (11)	14.80 (11)
selection terms	(0.000)			(0.421)	(0.192)

Notes. 1. $N = 550$; 2. GMM results obtained using the estimator 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) and (5), we also report corrected standard errors following Terza (2016). 6. The test of significance of the selection terms is a Wald test. Degrees of freedom and level of significance are in parentheses.

studied in this paper. SW's proposal is suitable for balanced panels and after making very particular assumptions about 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 presented in this paper.

5.2 Estimating models of tobacco consumption

The previous application is done on a small sample size in the cross-section dimension of $N = 550$, similar to the case with $N = 500$ in our Monte Carlo exercise. In this second application, we use a much larger sample size with $N = 2500$ and larger than the threshold where the difference between the AB and system GMM estimates converges to zero (see Figure C1). In more detail, we use the data in Jones and Labeaga (2003), who estimated rational addiction models of tobacco consumption. We make use of the repeated observations on tobacco expenditure in the ECPF from the third quarter of 1986 to the fourth of 1994. This is a rotating panel survey conducted by the

Spanish Statistical Office. Each quarter 3,200 individuals were interviewed, with replacement at a rate of 12.5 percent. Consequently, the maximum number of periods that an individual remains in the survey is eight and as an initial sample we use the balanced panel. The original size is 48,800 observations $N = 6100$ and $T = 8$. We follow Jones and Labeaga (2003) in using sample separation information to exclude those households who do not purchase tobacco in any of the eight observed periods since it does not induce endogenous selection. It implies dropping non-smokers ($N = 1957$). Those households who report zero and positive purchases may be affected by selection reflecting an intermittent sequence of quits and take-ups from smoking.

In this subsample Jones and Labeaga (2003) checked for some common pattern for the zeros in the smoking households' sample, but they did not find evidence either of corner solutions or clear sequences of starters-quitters. In these circumstances, they assume that the models underlying the zeros are type I Tobit specifications (i.e., zeros correspond to corner solutions). They estimated reduced form Tobit models, assuming normality, and to reduce the influence of distributional assumptions they adopted a semiparametric approach and estimated each of the T cross-section equations using Powell's (1986) Symmetrically Censored Least Squares (SCLS). SCLS is designed to accommodate standard Tobit-type censoring. The final model with and without correction is estimated with a sample of $N = 4041$ ($NT = 22520$), out of which 52 percent report eight positive purchases (see Table I in Jones and Labeaga, 2003 for further details).

We do not try to compare our results with Jones and Labeaga (2003), but we only like to compare the performance of our methods with a much larger sample size than in the previous application. In this sense, we are only interested in myopic models where only the lag of consumption, the price of tobacco, some time-varying demographics and time dummies enter the outcome equation (attending theoretical reasons, the price of tobacco does not enter the selection equation and can be used as an additional identification restriction). The results for the myopic model are presented in Table **6** (this is similar to a pure autoregressive model in the sense that the price of tobacco is an exogenous variable not included in the decision to start-quit smoking). The first column in Table **6** presents first-differenced AB myopic estimates obtained using predictions under censoring as in Jones and Labeaga (2003). It is important to note that the results of Jones and Labeaga (2003) and the results in this paper are not directly comparable. Jones and Labeaga (2003) control for non-smokers and they estimate a rational addiction model compared to our myopic behavior model

that does not control for non-smokers.

The rest of the columns in the table report the same estimators reported in the results of our first application. Columns (2) and (3) present AB and system GMM estimates obtained in the selected sample, but when we do not correct the consumption equation. In columns (4) and (5) we present AB and system GMM coefficients using a year-by-year correction for the equations in first differences (AB) and for all equations (system GMM).

Table 6: **Estimates of myopic models of tobacco consumption**

	(1)	*(2)*	*(3)*	*(4)*	*(5)*
	Jones Labeaga	*No correction*	*No correction*	*year by year correction first dif eq*	*year by year correction all equations*
	GMM	*AB*	*system*	*AB*	*system*
Lag real	0.2049***	0.1010***	0.1274***	0.0874***	0.0903***
tobacco consumption	(0.0147)	(0.0263)	(0.0189)	(0.0235)	(0.0203)
Real price	-1.0041***	-1.5900***	-0.8497***	-0.8828***	-0.6409***
of tobacco	(0.0672)	(0.3614)	(0.2278)	(0.3800)	(0.2267)
Observations	22520	22520	22520	22520	22520
Joint significance	–	–	–	39.88 (6)	183.61 (6)
selection terms				(0.000)	(0.000)

Notes. 1. $N = 4104$; 2. GMM results obtained in the sample of Jones and Labeaga (2003); 3. Quarter 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). The test of significance of the selection terms is a Wald test. Degrees of freedom and level of significance are in parentheses.

As usual in myopic models, we instrument lagged consumption using previous lags of consumption. However, qualitatively, the results for the autoregressive coefficient appear to reproduce the same characteristics found in the Monte Carlo exercises. We do not find big differences between the AB uncorrected and AB corrected for selection estimates. The interval of the mean plus / minus two standard errors contains both estimates with very high confidence. The same result occurs when comparing the system GMM uncorrected and corrected for selection coefficients. These results seem to suggest little need to correct the model, as suggested both by our theoretical and simulation results.

Finally, all the tests detect strong selectivity, but again correction does not seem to affect the

estimate of the lag, see columns (2) to (5). In this sense, the estimate of the lag in the model using the predicted latent variables in column (1) reports the highest difference, as expected, since the assumption is that zero purchases are due to censoring, i.e., they are corner solutions (see Jones and Labeaga, 2003). When the sample size in the individual dimension is sufficiently large, the AB and system GMM estimates are rather similar. This is true whether we correct the outcome equation for sample selection or not. Again, this is in line with our Monte Carlo results.

6 Concluding remarks

This paper studied the bias and consistency of classical panel data estimators including FE, RE and GMM estimators for both static and dynamic panel data models subject to potentially endogenous sample selection. We show that *a la Heckman* sample selection corrections are only needed when both equations have common covariates. In models without common covariates (and uncorrelated), regardless of the severity and even the complexity of the selection process (either with contemporaneous correlation only or with serial cross-correlation), standard estimators for the static model and the Arellano and Bond (1991) and the Anderson and Hsiao (1982) estimators for the dynamic model are consistent. Alternatively, the system GMM estimator is moderately biased regardless of the sample size. The bias is caused by the level orthogonality restrictions of the levels equations only, thereby implying that to correct the estimator we only need to correct those equations and not the equations in first differences. In the case the source of the bias is the correlation between the individual heterogeneous components in the outcome and selection equations, a simple control approach can handle this bias correction.

Alternatively, when the outcome and the selection equation have common covariates, we show the validity of simple corrections based on Wooldridge (1995), Rochina-Barrachina (1999) and Jiménez-Martín *et al.* (2009). When the errors in both equations are not serially cross-correlated we extend the proposal of Wooldridge (1995) to more complex cases, such as static models estimated in first differences or dynamic models. Alternatively, when they are serially cross-correlated ($cov(\varepsilon_{it}, u_{is} \neq 0); s < t$), we suggest using multivariate corrections.

We evaluate the finite sample performance of the classical panel data as well as GMM estimators in a Monte Carlo exercise. The results of our experiments confirm the theoretical predictions under

a variety of assumptions. Since sample size is crucial for the properties of the estimators and for the magnitude of the bias, we illustrate the properties of the estimators in two empirical applications differing in the number of individuals observed each period. The first one ($N = 550$) estimating female earnings equations using PSID data, and the second one ($N = 2500$) estimating myopic tobacco consumption equations using Spanish data. Our empirical studies give results in line with the results of the Monte Carlo study.

To conclude, as it is well known if the errors of the selection and outcome equations are not correlated, sample selection is not needed even if the two equations have common covariates. Moreover, the presence of common covariates also appear as a key determinant of the necessity of sample selection corrections *a la Heckman*. We believe that our findings could be of particular relevance for practitioners in situations where there are exclusion restrictions (implied by the theoretical model) or in increasingly common empirical studies based on experimental or quasi-experimental designs where the researcher have the control of factors influencing various stages of the experiment.

Acknowledgements

We are grateful to the Spanish Ministry of Economy for financial support through the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S) and projects ECO2017-83668-R and PID2020-114231RB-I00. We are specially grateful to María Rochina-Barrachina, for her contributions to an earlier draft of this work. We are also grateful to Manuel Arellano, Richard Blundell, Aureo de Paula, David Prieto, Juan M. Rodriguez-Poo, Martin Weidner and Frank Windmeijer for their useful comments. Also, the seminar audiences at UPF, UCL and Aarhus, the participants at the 2019, 2021 Panel Data Conferences, and the 2015 and 2019 IAAE conferences. Finally, the authors thank two referees and an editor for useful suggestions that helped improve the manuscript. All remaining errors are our responsibility.

References

[1] Anderson, T. W. Hsiao, C. (1982). 'Formulation and estimation of dynamic models using panel data', *Journal of Econometrics*, 18, 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*, 58, 277-297.

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

[4] 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, 23-48.

[5] Baltagi, B. (2021). *Econometric Analysis of Panel Data*, 6th edition, Springer, Switzerland.

[6] Becker, G. S. and Murphy, K. M. (1988). 'A theory of rational addiction', *Journal of Political Economy*, 96 675-700.

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

[8] Bover, O. and Arellano, M. (1997). 'Estimating limited-dependent variable models from panel data', *Investigaciones Economicas*, 21, 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. (1984). 'Panel data', in Z. Griliches, M. Intriligator (eds.), *Handbook of Econometrics*, Vol. 2, North-Holland, Amsterdam, Netherlands, 759-798.

[11] 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*, 10, 263–293.

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

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

[14] Han C. and Lee, G. (2022). 'Bias correction for within-group estimation of panel data models with fixed effects and sample selection', *Economics Letters*, 220. https://doi.org/10.1016/j.econlet.2022.110882

[15] Han C. and and Phillips P (2010), " GMM Estimation for dynamic Panels with fixed effects and strong intruments at unity", *Econometric Theory*, 2010, vol. 26, issue 1, 119-151.

[16] Hansen, L.P. (1982). 'Large sample properties of generalized method of moments estimators', *Econometrica*, 54, 1029-1054.

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

[18] Hausman, J. A. and Taylor, W. E. (1981). "Panel Data and Unobservable Individual Effects", *Econometrica*, 49(6), 1377–1398.

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

[20] Heckman, J. J. (1979). 'Sample Bias As A Specification Error', *Econometrica*, 47, 153-162.

[21] Horrace, W. C. and Oaxaca, R. L. (2006). 'Result on the bias and inconsistency of ordinary least squares for the linear probability model', *Economics Letters*, 90, 321-327.

[22] Jiménez-Martín, S. (1999). 'Controlling for endogeneity of strike variables in the estimation of wage settlement equations', *Journal of Labor Economics*, 17, 585-606.

[23] Jiménez Martín, S. (2006). 'Strike outcomes and wage settlements', *Labour*, 20, 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. (2003). 'Individual heterogeneity and censoring in panel data estimates of tobacco expenditures', *Journal of Applied Econometrics*, 18, 157-177.

[26] Knoef, M. and Been, J. (2015) 'Estimating a panel data sample selection model with part-time employment: Selection issues in wages over the life-cycle', WP, University of Leiden.

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

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

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

[30] Lai, H. P. and Tsay, W. J. (2018). 'Maximum simulation likelihood estimation of the panel data sample selection model', *Econometric Reviews*, 37, 744-759.

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

[32] Nickell, S. (1981) "Biases in Dynamic Models with Fixed Effects, *Econometrica*, Vol. 49, No. 6 (Nov., 1981), pp. 1417-1426.

[33] Olsen, R. J. (1980). 'A least squares correction for selectivity bias', *Econometrica*, 48, 1815-1820.

[34] Powell, J. (1986). 'Symmetrically trimmed least squares estimates for Tobit models', *Econometrica*, 54, 1435-1460.

[35] Raymond, W., Mohnen, P., Palm, F. and van der Loeff, S. S. (2007), 'The behavior of the maximum likelihood estimator of dynamic panel data sample selection models', CESIFO.

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

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

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

[39] Sargan, J. D. (1988). 'Testing for misspecification after estimating using instrumental variables', in: E. Maasoumi, ed., *Contributions to Econometrics: John Denis Sargan*, Vol. 1 (Cambridge University Press, Cambridge).

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

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

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

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

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

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

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

[47] Windmeijer, F. (2005), 'A finite sample correction for the variance of linear efficient two-step GMM estimators', *Journal of Econometrics*, 126, 25-51.

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

Appendices

A Consistency of the estimators when $\delta = 0$ and $x \perp z$

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 Sample selection corrections for IV estimators when $\delta \neq 0$ and $cov(\varepsilon_{it}, u_{is} \neq 0; s \leq t)$

In this section we develop the required correction for dynamic models in which IV is strictly necessary. For static model corrections see Wooldridge (1995) for the RE case and Rochina-Barrachina (1999) for the FD case.

B.1 Recap of a dynamic model

Consider an outcome variable y^*, which is related to its lagged value, and other variables included in the vector x.

$$y_{it}^* = \rho y_{it-1}^* + x_{it}\beta + \alpha_i + \varepsilon_{it} \qquad for \quad t_i \quad s.t. \quad d_{it} = 1; \tag{A1}$$

where d is the selection variable and α_i is an individual heterogeneity component independent of ε_{it}, the error term. ρ, β are parameters. x can be correlated with both the individual heterogeneity component and the error term. In addition we define $\omega_{it} = \alpha_i + \varepsilon_{it}$. Finally, note that when $\rho = 0$ we get the static model.
The observability of y^* is driven by the model for d, which is given by

$$d_{it}^* = z_{it}\gamma + x_{it}\delta + \eta_i + u_{it} = w_{it}\pi + \eta_i + u_{it}; \qquad d_{it} = 1\left[d_{it}^* \geq 0\right] \tag{A2}$$

where w (which combines z and x, being $x \perp z$) is a vector of strictly exogenous regressors (with respect to u once we allow for w to be correlated with η_i), η_i is a term capturing unobserved individual heterogeneity and u_{it} is an error term. Assumptions about the components of (A1) and (A2) will be given in the next subsections.

Furthermore, in general, $\eta_i + u_{it}$ and $\alpha_i + \varepsilon_{it}$ can be serially cross-correlated, that is $cov(\varepsilon_{it}, u_{is}) \neq 0; s \leq t$.

B.2 General assumptions for the selection equation

•***B1:*** *The conditional expectation of η_i given $\bar{w}_i$ is linear.*
Following Mundlak (1978), it is assumed that the conditional expectation of the individual effects in the selection equation is linear in the time means of all exogenous variables (alternatively, we can also use Chamberlain's, 1984, approach): $\eta_i = \bar{w}_i\theta + c_i$, where c_i is a random component independent of w_i (recall that w represents the combination of z and x).
•***B2:*** *The errors in the selection equation, $\nu_{it} = u_{it} + c_i$, are independent of w_i and normal $\left(0, \sigma_t^2\right)$.*
Under ***B1*** and ***B2*** the reduced form selection rule of (A2) is $d_{it}^* = w_{it}\pi + \bar{w}_i\theta + \nu_{it}$, $d_{it} = 1\left\{w_{it}\pi + \bar{w}_i\theta + \nu_{it} \geq 0\right\} = 1\left\{H_{it} + \nu_{it} \geq 0\right\}$.

The reduced form selection rule $d_{it}^* = w_{it}\pi_t + \bar{w}_i\theta_t + \nu_{it}$ is not only compatible with ***B1*** (to allow the w to be correlated with the individual effect in the selection equation) but also with a dynamic

model for the selection rule such as: $d^*_{it} = \rho_d d^*_{it-1} + w_{it}\pi_t + \eta_i + u_{it}$, where $d^*_{i0} = \bar{w}_i\pi_0 + u_{i0}$ (initial condition) and $\eta_i = \bar{w}_i\theta + c_i$ (as in ***B1***). In this case ν_{it} will be a function of $u_{i0}, ..., u_{it}, c_i$, but still independent of w_i.

B.3 Bias correction

B.3.1 Correction of the first differenced (FD) equations

Let us consider the first-differenced model:

$$\Delta y_{it} = \rho \cdot \Delta y_{it-1} + \Delta x_{it}\beta + \Delta\varepsilon_{it} \tag{A3}$$

We will need a sample of individuals with $d_{it} = d_{it-1} = d_{it-2} = 1$, and, therefore, in general the sample selection correction term will come from a trivariate probit:

$$\Delta y_{it} = \rho \cdot \Delta y_{it-1} + \Delta x_{it}\beta + E\left[\Delta\varepsilon_{it}\,|w_i, d_{it} = d_{it-1} = d_{it-2} = 1\right] + \Delta e_{it}. \tag{A4}$$

We follow Tallis (1961) to work it out: $E\left[\Delta\varepsilon_{it}\,|w_i, d_{it} = d_{it-1} = d_{it-2} = 1\right]$ under a 4-variant normal distribution assumption. In fact, by assuming a linear projection of the errors in the main equation $\Delta\varepsilon_{it}$ on the errors in the selection equations in t , $t-1$ and $t-2$, we do not need a 4-variant normal distribution for the errors in both equations $[\Delta\varepsilon_{it}, \nu_{it}, \nu_{it-1}, \nu_{it-2}]$, but only a trivariate normal distribution for the errors in the selection equation $(\nu_{it}, \nu_{it-1}, \nu_{it-2})$.

•***B3***: The errors $[\Delta\varepsilon_{it}, \nu_{it}, \nu_{it-1}, \nu_{it-2}]$ are 4-variate normally distributed and independent of w_i. Therefore,

$$\begin{aligned} &E\left[\Delta\varepsilon_{it}\,|w_i, d_{it} = d_{it-1} = d_{it-2} = 1\right] = \sigma_{\Delta\varepsilon_t, \frac{\nu_t}{\sigma_t}}\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) \\ &+\sigma_{\Delta\varepsilon_t, \frac{\nu_{t-1}}{\sigma_{t-1}}}\lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) + \sigma_{\Delta\varepsilon_t, \frac{\nu_{t-2}}{\sigma_{t-2}}}\lambda\left(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right). \end{aligned} \tag{A5}$$

where $H_{is} = w_{is}\pi - E(\eta_i|w_i)$ for $s = t, \quad t-1, \quad t-2$, and,

$$\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) = \frac{\phi(H_{it})\Phi_2\left(\left(H_{it-1} - \varrho_{t,t-1}H_{it}\right)\Big/\left(1-\varrho^2_{t,t-1}\right)^{1/2}, \left(H_{it-2} - \varrho_{t,t-2}H_{it}\right)\Big/\left(1-\varrho^2_{t,t-2}\right)^{1/2}, \varrho_{t-1,t-2.t}\right)}{\Phi_3(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})},$$

$$\lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) = \frac{\phi(H_{it-1})\Phi_2\left(\left(H_{it} - \varrho_{t,t-1}H_{it-1}\right)\Big/\left(1-\varrho^2_{t,t-1}\right)^{1/2}, \left(H_{it-2} - \varrho_{t-1,t-2}H_{it-1}\right)\Big/\left(1-\varrho^2_{t-1,t-2}\right)^{1/2}, \varrho_{t,t-2.t-1}\right)}{\Phi_3(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})},$$

$$\lambda\left(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) = \frac{\phi(H_{it-2})\Phi_2\left(\left(H_{it} - \varrho_{t,t-2}H_{it-2}\right)\Big/\left(1-\varrho^2_{t,t-2}\right)^{1/2}, \left(H_{it-1} - \varrho_{t-1,t-2}H_{it-2}\right)\Big/\left(1-\varrho^2_{t-1,t-2}\right)^{1/2}, \varrho_{t,t-1.t-2}\right)}{\Phi_3(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})}$$

where $\phi\left(\right)$ is the standard normal density function, and $\Phi_2\left(\right)$, $\Phi_3\left(\right)$ are the standard bivariate and trivariate normal cumulative distribution functions, respectively. The $\varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}$ are all the possible correlation coefficients between the errors in the selection equation in the three time periods.

To construct estimates of the $\lambda()$ terms, first, the coefficients in the *Hs* will be jointly determined with $\varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}$, using a trivariate probit for the three time periods. Doing this we will get a predicted value for the trivariate probability $\Phi_3(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})$ that appears in the denominator of the $\lambda()$ terms. Second, we will get also estimates for the two arguments of the type $(H_{is} - \varrho_{t,s}H_{it}) \Big/ \left(1-\varrho_{t,s}^2\right)^{1/2}$ in the bivariate probabilities $\Phi_2()$. Third, we will perform all the involved bivariate probabilities $\Phi_2()$ and estimate the partial correlation coefficients $\varrho_{t-1,t-2.t}$, $\varrho_{t,t-2.t-1}$, $\varrho_{t,t-1.t-2}$ for fixed $H_{it}, H_{it-1}, H_{it-2}$, respectively. Fourth, we will get a predicted value for the bivariate probabilities $\Phi_2()$ that are in the numerators of the $\lambda()$ terms multiplied by the corresponding $\phi(H_{is})$.

Under stationarity $\sigma_{\varepsilon_t, \frac{\nu_t}{\sigma_t}} = \sigma_{\varepsilon_{t-1}, \frac{\nu_{t-1}}{\sigma_{t-1}}}$, and we will call it σ_0. Now (A5) becomes:

$$
\begin{aligned}
&E[\Delta\varepsilon_{it} | w_i, d_{it} = d_{it-1} = d_{it-2} = 1] = \\
&\sigma_0 \{\lambda(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) - \lambda(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})\} \\
&-\sigma_{\varepsilon_{t-1}, \frac{\nu_{t,2}}{\sigma_t}} \lambda(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) + \sigma_{\varepsilon_t, \frac{\nu_{t-1}}{\sigma_{t-1}}} \lambda(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) \\
&+\sigma_{\varepsilon_t, \frac{\nu_{it-2}}{\sigma_{t-2}}} \lambda(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) - \sigma_{\varepsilon_{t-1}, \frac{\nu_{it-2}}{\sigma_{t-2}}} \lambda(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})
\end{aligned}
\tag{A6}
$$

In this equation the correlation $\sigma_{\varepsilon_{t-1}, \frac{\nu_t}{\sigma_t}}$ does not have to be equal to the correlations $\sigma_{\varepsilon_t, \frac{\nu_{t-1}}{\sigma_{t-1}}} = \sigma_{\varepsilon_{t-1}, \frac{\nu_{it-2}}{\sigma_{t-2}}}$, or $\sigma_{\varepsilon_t, \frac{\nu_{it-2}}{\sigma_{t-2}}}$, but let us call $\sigma_{\varepsilon_{t-1}, \frac{\nu_t}{\sigma_t}} = \sigma_{+1}$, $\sigma_{\varepsilon_t, \frac{\nu_{t-1}}{\sigma_{t-1}}} = \sigma_{\varepsilon_{t-1}, \frac{\nu_{it-2}}{\sigma_{t-2}}} = \sigma_{-1}$, and $\sigma_{\varepsilon_t, \frac{\nu_{it-2}}{\sigma_{t-2}}} = \sigma_{-2}$ under stationarity.

Then equation (A6) becomes:

$$
\begin{aligned}
&E[\Delta\varepsilon_{it} | w_i, d_{it} = d_{it-1} = d_{it-2} = 1] = \\
&\sigma_0 \lambda(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) - \sigma_0 \lambda(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) \\
&-\sigma_{+1} \lambda(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) + \sigma_{-1} \lambda(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) \\
&+\sigma_{-2} \lambda(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) - \sigma_{-1} \lambda(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) = \\
&(\sigma_0 - \sigma_{+1}) \lambda(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) - (\sigma_0 - \sigma_{-1}) \lambda(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) \\
&+ (\sigma_{-2} - \sigma_{-1}) \lambda(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})
\end{aligned}
\tag{A7}
$$

Further, if we assume an exchangeability condition like the one in Kyriazidou (1997), this implies $\sigma_{+1} = \sigma_{-1}$ (let us call them simply σ) and in this case equation (A7) becomes:

$$
\begin{aligned}
&E[\Delta\varepsilon_{it} | w_i, d_{it} = d_{it-1} = d_{it-2} = 1] = \\
&\bar{\sigma} \{\lambda(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) - \lambda(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})\} \\
&+\bar{\sigma}_{-2} \lambda(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})
\end{aligned}
\tag{A8}
$$

where $\bar{\sigma} = \sigma_0 - \sigma$ and $\bar{\sigma}_{-2} = \sigma_{-2} - \sigma$. That means that correcting for sample selection with longitudinal correlation of the errors increases the dimension of regressors by two.

Importantly, when there is no serial cross-correlation between the errors in the outcome and the selection equation, $\varrho_{t,t-1} = \varrho_{t,t-2} = \varrho_{t-1,t-2} = 0$, also $\varrho_{t-1,t-2,t} = \varrho_{t,t-2,t-1} = \varrho_{t,t-1,t-2} = 0$, and we have that

$$
\lambda(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) = \phi(H_{it}) / \Phi(H_{it}) = \lambda(H_{it}),
$$

$$\lambda\left(H_{it-1},H_{it},H_{it-2},\varrho_{t,t-1},\varrho_{t,t-2},\varrho_{t-1,t-2}\right)=\phi\left(H_{it-1}\right)/\Phi\left(H_{it-1}\right)=\lambda\left(H_{it-1}\right),$$

$$\lambda\left(H_{it-2},H_{it},H_{it-1},\varrho_{t,t-1},\varrho_{t,t-2},\varrho_{t-1,t-2}\right)=\phi\left(H_{it-2}\right)/\Phi\left(H_{it-2}\right)=\lambda\left(H_{it-2}\right),$$

The corrected outcome equation (A5) becomes:

$$E\left[\Delta\varepsilon_{it}\left|w_i,d_{it}=d_{it-1}=d_{it-2}=1\right.\right]=\sigma_{(\varepsilon_t),\frac{\nu_t}{\sigma_t}}\lambda\left(H_{it}\right)-\sigma_{(\varepsilon_{t-1}),\frac{\nu_{t-1}}{\sigma_{t-1}}}\lambda\left(H_{it-1}\right) \tag{A9}$$

and the model has to include as new regressors correcting for sample selection the standard Heckman lambda terms coming from univariate probits in *t* and *t-1*. Under stationarity (A9) becomes $\sigma_0\left\{\lambda\left(H_{it}\right)-\lambda\left(H_{it-1}\right)\right\}$.

B.3.2 Correction of the level equations

Let us consider now the estimation of the levels equations.

$$\begin{array}{l} y_{it}=\rho y_{it-1}+x_{it}\beta+\bar{z}_i\psi+E\left[\omega_{it}\left|z_i,d_{it}=d_{it-1}=d_{it-2}=1\right.\right]+e_{it}= \\ \rho y_{it-1}+x_{it}\beta+\bar{z}_i\psi+\sigma_{\omega_t,\frac{\nu_t}{\sigma_t}}\lambda\left(H_{it},H_{it-1},\varrho_{t,t-1}\right)+\sigma_{\omega_t,\frac{\nu_{t-1}}{\sigma_{t-1}}}\lambda\left(H_{it-1},H_{it},\varrho_{t,t-1}\right)+e_{it} \end{array}, \tag{A10}$$

and consider the following assumption:

•***B3':*** *The errors* $[\omega_{it},\nu_{it},\nu_{it-1}]$ *are trivariate normally distributed and independent of* z_i.

Under stationarity $\sigma_{\omega_t,\frac{\nu_t}{\sigma_t}}=\sigma_0$ and $\sigma_{\omega_t,\frac{\nu_{t-1}}{\sigma_{t-1}}}=\sigma_{-1}$, and (A10) becomes:

$$y_{it}=\rho y_{it-1}+x_{it}\beta+\bar{w}_i\psi+\sigma_0\lambda\left(H_{it},H_{it-1},\varrho_{t,t-1}\right)+\sigma_{-1}\lambda\left(H_{it-1},H_{it},\varrho_{t,t-1}\right)+e_{it} \tag{A11}$$

To construct estimates of the $\lambda()$ terms the coefficients in the *Hs* will be jointly determined with $\varrho_{t,t-1}$, using a bivariate probit for each pair of time periods.

Importantly, when the errors in the outcome and selection equations are not time-series correlated $\varrho_{t,t-1}=0$, then $\sigma_{-1}=0$, and (A10) becomes:

$$\begin{array}{l} y_{it}=\rho y_{it-1}+x_{it}\beta+\bar{w}_i\psi+E\left[\nu_{it}\left|z_i,d_{it}=1\right.\right]+e_{it}= \\ \rho y_{it-1}+x_{it}\beta+\bar{w}_i\psi+\sigma_0\lambda\left(H_{it}\right)+e_{it} \end{array} \tag{A12}$$

and we come back to univariate probits per each *t*.

B.4 Summary and empirical guidelines

When the errors in the outcome and selection equations are (cross) serially correlated (that is, when $cov(\varepsilon_{it},u_{is}\neq 0;s<t)$ we generally require sample selection correction terms that require estimation of a trivariate probit and we need at least 3 periods per individual. For the differences equation estimation, the relevant samples are constructed by picking up at least three consecutive treatment outcomes or alternatively three non-treatment outcomes per individual. When after selecting the observations in this way the treatment sample is not large enough to allow the identification of the relevant parameters of the equation, we estimate this equation by levels estimation exploiting only the extra moment conditions of system GMM (Arellano and Bover, 1995; Blundell and Bond,

1998) *versus* GMM (Arellano and Bond, 1991). In the latter case we require samples with two consecutive outcomes of the same regime.

B.4.1 Using standard software

In the first differences model, under the assumption that $cov(\varepsilon_{it}, u_{is} = 0; s < t)$ and assuming stationarity, (A9) can be estimated with the Stata *xtabond* command. In the more general stationary only case, (A8) can be estimated with a modified version of the *xtabond* command adding two regressors: $\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) - \lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right)$; and
$\lambda(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})$.

With System-GMM estimation, and under stationarity only, joint estimation of equations (A12) and (A9) with the *xtdpdsys* Stata System-GMM command if we restrict the level sample in the same way as the first differenced one. However, the Stata command have to be adapted to allow for different coefficients of the sample selection correction terms in the equation in levels ($\sigma_{\omega_t, \frac{\nu_t}{\sigma_t}}$ in (A12)) than in the equation in time differences ($\sigma_{\varepsilon_t, \frac{\nu_t}{\sigma_t}}$ in (A9)).
Under *Simplification 1,* it will be more difficult to adapt standard software because, in addition to adding different regressors to the levels ($\{\lambda\left(H_{it}, H_{it-1}, \varrho_{t,t-1}\right), \lambda\left(H_{it-1}, H_{it}, \varrho_{t,t-1}\right)\}$) and the differenced equations ($\{\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) - \lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right)\}$, $\lambda\left(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right)$, we have to allow for different parameters associated with the sample selection correction terms in the level and differenced equations.

B.5 Semiparametric model estimation

B.5.1 Correction of level equations

Consider the level model:

$$y_{it} = \rho y_{it-1} + x_{it}\beta + \bar{z}_i\psi + E\left[\omega_{it} \left| w_i, d_{it} = d_{it-1} = 1\right.\right] + e_{it}$$

,
where the conditional mean is now an unknown function of the selection indices H_{it}, H_{it-1}, that is:

$$E\left[\omega_{it} \left| w_i, d_{it} = d_{it-1} = j\right.\right] = \varphi_{jt,t-1}\left(H_{it}, H_{it-1}\right) = \varphi_{jit,t-1}$$

Errors can depend on the w_i only through these indices (what is called a "double index" assumption). Now (A10) becomes

$$y_{it} = \rho y_{it-1} + x_{it}\beta + \bar{z}_i\psi + \varphi_{jit,t-1} + e_{it}$$

Once selection indices size has been obtained, in a first stage, by a normal, logistic or the Heckman's lambda (inverse Mill's ratio) transformation, the unknown function $\varphi_{jit,t-1}$ is approximated nonparametrically by a polynomial of degree q on the transformation of the indices H_{it}, H_{it-1}. In the

general case of absence of stationarity, we will interact the terms of the polynomial with time-pair dummies.

B.5.2 Correction of first differenced equations

Consider the first differenced model:

$$\Delta y_{it} = \rho \Delta y_{it-1} + \Delta x_{it}\beta + E\left[\Delta \varepsilon_{it} \left| w_i, d_{it} = d_{it-1} = d_{it-2} = j \right.\right] + \Delta e_{it}$$

where instead of giving a parametric expression for $E\left[\Delta \varepsilon_{it} \left| w_i, d_{it} = d_{it-1} = d_{it-2} = j \right.\right]$ in (A5) we write $E\left[\Delta \varepsilon_{it} \left| w_i, d_{it} = d_{it-1} = d_{it-2} = j \right.\right] = \varphi_{jt,t-1,t-2}\left(H_{it}, H_{it-1}, H_{it-2}\right) = \varphi_{jit,t-1,t-2}$, where the conditional mean is now an unknown function of the selection indices $H_{it}, H_{it-1}, H_{it-2}$. Errors can depend on the w_i only through these indices (what is called a "triple index" assumption).
Now (A4) becomes $\Delta y_{it} = \rho \Delta y_{it-1} + \Delta x_{it}\beta + \varphi_{jit,t-1,t-2} + \Delta e_{it}$. The unknown function $\varphi_{jit,t-1,t-2}$ is approximated non-parametrically by a polynomial of degree q on the transformation of the indices $H_{it}, H_{it-1}, H_{it-2}$. Note that in the general case of absence of stationarity, we will interact the terms of the polynomial with time-triples dummies. The identification of the selection indexes, or first step, can be achieved by assuming a normal or logístic transformation. We could estimate the first step by using a semiparametric method for binary choice with panel data.

Besides the (parametric or semi-parametric) specification of the sample selection correction terms, the models will be finally estimated by GMM (AB) or system-GMM (when $\rho \neq 0$ and/or x is endogenous) or RE,FE,FD (when $\rho = 0$ and x is exogenous).

C Additional tables and figures

Table C1. 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%.

Table C2. Average bias, RMSE and coverage rates of C.I. in the static model. x strictly exogenous. T=7; 500 replications

x in selection	Corrected	FE estimator av. bias	FE estimator RMSE	FE estimator CICR[2]	FE estimator ERF sel. term	FD estimator av. bias	FD estimator RMSE	FD estimator CICR	FD estimator ERF sel. term	RE (GLS) estimator $cov((z_i,x_i),\alpha_i)=0$, $cov(x_i,\alpha_i)=0$ av. bias	RE (GLS) estimator RMSE	RE (GLS) estimator CICR	RE (GLS) estimator ERF sel. term	Corrected RE (GLS) estimator[1] $cov((z_i,x_i),\alpha_i)\neq 0$,$cov(x_i,\alpha_i)\neq 0$ av. bias	Corrected RE (GLS) estimator RMSE	Corrected RE (GLS) estimator CICR	Corrected RE (GLS) estimator ERF sel. term
Panel A: N = 500; endogenous selection, $cov(x,z)=0$																	
No	No	0.0003	0.0214	0.936		-0.0007	0.0301	0.958		0.0005	0.0160	0.952		-0.0001	0.0218	0.962	
Yes	No	-0.0474	0.0536	0.726		-0.0427	0.0554	0.558		-0.0642	0.0671	0.094		-0.0520	0.0583	0.482	
Yes	Yes[3]	0.0011	0.0295	0.926	0.950	0.0010	0.0388	0.954	0.832	-0.0034	0.0245	0.938	0.998	0.0007	0.0297	0.95	0.996
Panel A: N = 500; exogenous selection, $cov(x,z)=0$																	
Yes	Yes[3]	0.0010	0.0261	0.924	0.050	0.0008	0.0349	0.944	0.098	0.0009	0.0218	0.944	0.070	0.0045	0.0294	0.942	0.060
Panel B: N = 5000; endogenous selection, $cov(x,z)=0$																	
No	No	-0.0001	0.0070	0.948		0.0002	0.0093	0.956		0.0001	0.0052	0.942		-0.0000	0.0072	0.954	
Yes	No	-0.0469	0.0476	0.022		-0.0413	0.0427	0.000		-0.0640	0.0642	0.000		-0.0512	0.0519	0.000	
Yes	Yes[3]	0.0020	0.0097	0.930	1.000	0.0023	0.0128	0.942	1.000	-0.0024	0.0081	0.930	1.000	0.0031	0.0103	0.914	1.000
Panel B: N = 5000; exogenous selection, $cov(x,z)=0$																	
Yes	Yes[3]	0.0001	0.0084	0.914	0.038	0.0005	0.0113	0.940	0.086	0.0001	0.0067	0.966	0.038	0.0041	0.0104	0.906	0.12
Panel C: N = 500; endogenous selection, $cov(x,z)\neq 0$																	
No	No	-0.0253	0.0305	0.654		-0.0315	0.0399	0.628		-0.0211	0.0251	0.656		-0.0010	0.0162	0.958	
No	Yes[4]	-0.0011	0.0177	0.918	0.978	-0.0012	0.0260	0.944	0.926	-0.0008	0.0142	0.946	0.998	0.0009	0.0163	0.944	0.986
Panel C: N = 5000; endogenous selection, $cov(x,z)\neq 0$																	
No	No	-0.0253	0.0259	0.002		-0.0315	0.0324	0.002		-0.0212	0.0216	0.002		-0.0001	0.0052	0.954	
No	Yes[4]	-0.0004	0.0055	0.916	1.000	0.0002	0.0082	0.948	1.000	-0.0004	0.0045	0.944	1.000	0.0017	0.0055	0.924	1.000

1. In the last column (corrected GLS estimator) we add the mean of z and x as covariates in both the selection and the outcome equation to control for the correlation between z, x and η, α.
2. CICR. 95 % Confidence intervals coverage rates, ie. CICR= $\sum_S \mathbf{1}(\hat{a} - 1.96 * s.e.(\hat{a}) < a < \hat{a} + 1.96 * s.e.(\hat{a}))/S$, where S is the number of simulations, a is the true parameter and $\hat{a}$ is the estimate of a in each simulation.
3. In Panels A and B the correction is obtained from a year by year probit with z, x as covariates.
4. In Panel C the correction is $E(x|z)$, $cov(x, \Psi) \neq 0$ where $\Psi \in z$.

Table C3. Average bias, RMSE and coverage rates of C.I. in the purely AR(1) model. T=7; 500 replications. GMM-IV and Han and Phillips estimators

GMM-IV estimators. Results using the first seven generated observations

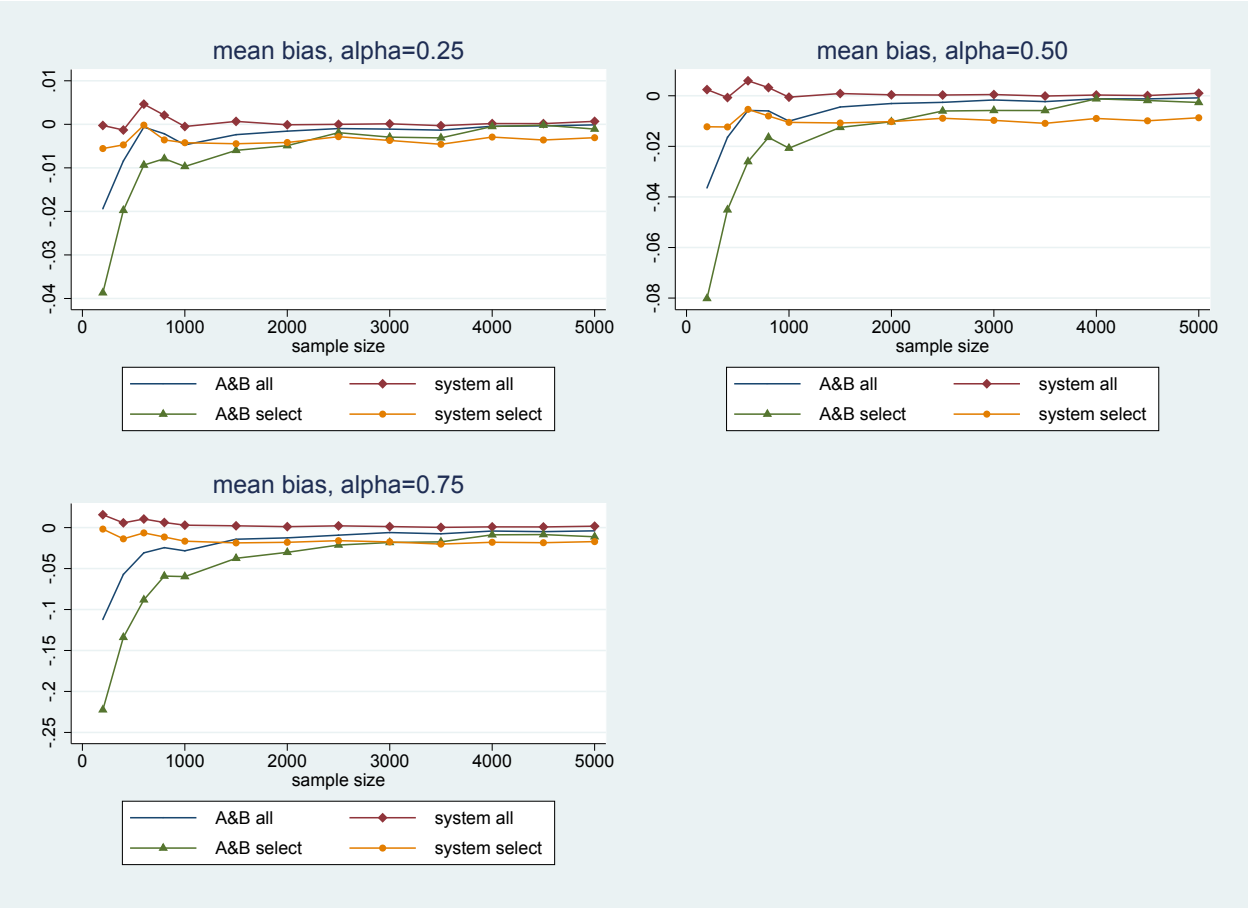
	Estimates with the full sample						Estimates with the selected sample					
	AB estimator			system			AB estimator			system		
ρ	av. bias	RMSE	CICR[1]	av. bias	RMSE	CICR	av. bias	RMSE	CICR	av. bias	RMSE	CICR
Panel A: N=500												
0.25	-0.0054	0.0435	0.930	0.0012	0.0322	0.940	-0.0168	0.7687	0.940	-0.0047	0.7559	0.936
0.50	-0.0135	0.0591	0.916	0.0018	0.0386	0.928	-0.0275	0.5317	0.934	-0.0077	0.5101	0.932
0.75	-0.0102	0.0388	0.936	0.0017	0.0265	0.954	-0.0212	0.2769	0.944	0.0011	0.2510	0.952
Panel B: N=5000												
0.25	-0.0007	0.0130	0.946	-0.0006	0.0096	0.942	-0.0021	0.7523	0.956	-0.0047	0.7548	0.928
0.50	-0.0017	0.0163	0.936	-0.0009	0.0111	0.936	-0.0036	0.5040	0.940	-0.0086	0.5088	0.900
0.75	-0.0010	0.0105	0.970	-0.0002	0.0080	0.950	-0.0029	0.2535	0.944	-0.0009	0.2511	0.948

Han and Phillips LS estimators.
T=7 discarding the initial 13 observations

	Estimates with the full sample			Estimates with the selected sample		
	HP LS estimator			HP LS estimator		
ρ	av. bias	RMSE	CICR	av. bias	RMSE	CICR
Panel A: N=500						
0.25	-0.0005	0.7513	0.944	0.0007	0.7504	0.960
0.50	-0.0013	0.5025	0.950	0.0023	0.4995	0.962
0.75	-0.0005	0.2530	0.952	0.0061	0.2479	0.956
Panel B: N=5000						
0.25	-0.0009	0.7509	0.960	0.0007	0.7494	0.946
0.50	-0.0007	0.5008	0.960	0.0035	0.4967	0.950
0.75	-0.0003	0.2505	0.956	0.0077	0.2428	0.928
T=7 using the first seven observations						
Panel A: N=500						
0.25	0.0651	0.6857	0.512	0.0639	0.6872	0.644
0.50	0.1777	0.3242	0.004	0.1789	0.3239	0.016
0.75	0.4293	0.1835	0.000	0.4352	0.1909	0.000
Panel B: N=5000						
0.25	0.0630	0.6871	0.002	0.0638	0.6863	0.002
0.50	0.1756	0.3246	0.000	0.1793	0.3210	0.000
0.75	0.4268	0.1774	0.000	0.4347	0.1854	0.000

1. CICR. 95 % Confidence intervals coverage rates, ie. CICR= $\sum_S \mathbf{1}(\hat{a} - 1.96 * s.e.(\hat{a}) < a < \hat{a} + 1.96 * s.e.(\hat{a}))/S$, where S is the number of simulations, a is the true parameter and $\hat{a}$ is the estimate of a in each simulation.

Figure C1. Average bias of the AB and system estimators in the full sample (NT observations) and the endogenously selected sample



Notes.

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

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

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

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

Table C4. AR(1) log hourly earnings equation

	(1)	*(2)*	*(3)*	*(4)*
	2SLS-IV	*No correction*	*No correction*	*year by year correction of lev eq. only*
		AB	*system*	*system*
Lag log hourly earnings	0.1522**	0.1029**	0.1798***	0.2354***
	(0.0489)	(0.0377)	(0.0434)	(0.0444)
Observations	5033	5033	5033	5033
Joint significance selection terms				105.13 (11) (0.000)

Notes: 1. $N = 550$; 2. Annual dummies are included in all specifications; 3. *** significant at 1%; ** significant at 5%; * significant at 10%; 4. The standard errors have been corrected following Windmeijer (2005). In column (4), we also report corrected standard errors following Terza (2016) 5. The test of significance of the selection terms is a Wald test. Degrees of freedom and level of significance are in parentheses.