Documento de Trabajo - 2020/06

Consistent estimation of panel data sample selection models

Sergi Jiménez-Martín
(Universitat Pompeu Fabra, BGSE and FEDEA)

José M. Labeaga
(UNED)

Majid al Sadoon
(Durham University Business School)

fedea

Consistent estimation of panel data sample selection models *

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

April 2020

Abstract

We analyse the properties of classical (fixed effect, first-differences and random effects) as well as generalised method of moments-instrumental variables estimators in either static or dynamic panel data sample selection models. We show that the correlation of the unobserved errors is not sufficient for non-consistency to arise, but the presence of common (and/or non-independent) non-deterministic covariates in the selection and outcome equations is generally necessary. When both equations do not have covariates in common and independent of each other, we show the consistency of fixed effects and random effects estimators in static models with exogenous covariates. Furthermore, the first-differenced generalised method of moments estimator uncorrected for sample selection of Arellano and Bond (1991) as well as the instrumental variables estimator of Anderson and Hsiao (1982) are consistent for autorregressive models even with endogenous covariates. The same results hold when theboth equations have no covariates in common but they are correlated, once we account for such correlation. Under the same circumstances, the system generalised method of moments estimator (Arellano and Bover, 1995, and Blundell and Bond, 1998) has a moderate bias. Alternatively, when both equations have covariates in common we suggest the appropriate correction method, being the serial correlation of the errors a key determinant of the choice. The finite sample properties of the proposed estimators and solutions are evaluated using a Monte Carlo study. We also do two different applications to log earning equations for females using the Panel Study of Income Dynamics and to tobacco consumption models using the Spanish Continuous Family Expenditure Survey.

JEL Codes: J52, C23, C24

Keywords: Panel data, sample selection, static and dynamic models, generalized method of moments

*We are grateful to the Spanish Ministry of Economy for financial support through projects ECO2014-52238-R, and ECO2017-83668-R. We are specially grateful with María Rochina-Barrachina, who contributed to parts of this work. We are also grateful to Manuel Arellano, Badi Baltagi, Richard Blundell, Aureo de Paula, David Prieto, Juan M. Rodriguez-Poo, Martin Weidner and Frank Windmeijer for some very useful comments, to the seminar audience at UPF and UCL, to participants at the 2019 Panel Data Conference in Vilnius, and the 2015 and 2019 IIAE conferences. All remaining errors are our responsibility. The usual disclaimer applies.

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

‡UNED

§Durham University Business School

1 Introduction

The problems of self-selection, non-response and attrition are common in datasets containing economic variables. Their presence generate manageable models in cross-sections. However, correlated heterogeneity together with endogenous attrition, non-response or selection complicate the models with unbalanced panel data (Baltagi, 2013). 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 likely to be used more frequently in the future. In this context, we believe that it is important to highlight advantages and problems in the performance of different estimators and to draw researchers' attention to potential pitfalls in using them in empirical studies.

In this paper we focus on the estimation of panel data sample selection models. We consider a variety of cases for the outcome of interest and a simple form, easily generalizable, for the selection equation. The error components of both equations can be correlated with a very general correlation structure. Departing from the simplest situation, we present an exercise including all important features in the model one by one to test their individual and joint effects on the bias of some of the classical estimators (fixed effects -FE-, random effects -RE-, or first differences -FD-) as well as the more sophisticated generalized method of moments (GMM) estimators.

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

The first two cases are far less common that the others. They typically involve sample selection related to involuntary factors, not linked with the individual or firm characteristics (being the ongoing Covid-19 crisis an excellent example). They imply the determinants of the intensive (the observability rule) and the extensive (the outcome equation) margins are totally different. Yet the unobserved components (time and time-invariant) can be correlated causing endogenous sample selection. However, as we shall illustrate latter on, under these circumstances sample selection corrections are not necessary to obtain consistent estimates of the parameters of interest. Alternatively, in the last two, most common cases, sample selection correction (a la Heckman) will be necessary, and, more importantly, will have increasing complexity.

For case (a) we distinguish static and dynamic sample selection models. In the static model without common (and independent) covariates between the outcome and the selection equation (let us call them x and z respectively), we show that all the classical (Fixed Effects, First Differences and Random Effects GLS, the latter under the additional condition of no correlation of the covariates with the heterogeneity effect) and GMM (in case of endogeneity of any regressor) estimators are consistent.

Similarly to the case above, in dynamic models without common time-varying covariates (the

purely AR(1) as well as the Montecarlo study in Raymond et al (2007), lately applied in Raymond et al (2010), are paramounts examples of this case) the uncorrected for selection GMM estimator of Arellano and Bond (AB, 1991) as well as the less efficient of Anderson and Hsiao (AH, 1982), are consistent regardless of the exogenous or endogenous nature of the selection.[1] Furthermore, we show that the additional orthogonality restrictions implied by the system GMM estimator (Arellano and Bover, 1995; Blundell and Bond, 1998) are not valid under endogenous selection. However, the inconsistency of the system estimator is small and hardly induces bias, even and especially in small samples, when the time-invariant heterogeneity components in the outcome and selection equations are not correlated. All these also apply to models with exogenous, predetermined or endogenous covariates, which are, in turn, not present in the selection equation.

In models without common but correlated covariates (case (b)), we show that still there is no need to control for the correlation of the time varying errors, provided that, as we shall describe latter on, we control (instrument) in the outcome equation for relationship between the covariates in both equations. Our approach will be similar to the Olsen (1980) solution for the least squares model.

For case (c), that is for models with at least one common covariate[2], we propose an extension of Wooldridge (1995) and Rochina-Barrachina (1999) based in the estimation of year-by-year probits (although we also suggest some semiparametric alternatives). In static models in levels, we strictly follow the proposal of Wooldridge (1995) and we 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 can apply and extend Wooldridge's proposal safely.

Finally, for case (d), when both equations have covariates in common and the time varying errors are serially cross-correlated, we suggest a multivariate corrections. Interestingly, in models with predetermined or endogenous covariates the selection terms need to be instrumented accordingly.

Although is not the focus of this paper, testing between the alternative cases described above is not complicated at all. For example a simple t-test or Wald test allow to check the significance of x in the selection equation. In case it is not detected, a test of the $E(x|z)$ checks for the necessity of correcting for the correlation between x and z. Finally, to distinguish between (c) and (d) we can test the correlation between the time variant errors in the outcome and the lagged (one and twice if necessary) time variant 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 the two equations, our results suggest non-necessity of correcting the classical static estimators or the first-differences AB estimates in the selected sample. In models with covariates in common, we show that our suggestions for correction are able to eliminate or significantly reduce the selection bias.

[1] 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.

[2] The common covariate can be also the lagged outcome, as in Gayle and Viauroux (2007).

Our work contributes to the literature in several dimensions. First by showing the non-necessity to correct for selectivity (even with a high degree of correlation) when both equations do not have time-varying covariates in common. Second, by suggesting simple methods to correct the outcome equation when both equations have common covariates. Combining these contributions, we conclude that the key determinant of the necessity of sample selection correction *a la Heckman* is the presence of common covariates and not whether the errors of the selection and outcome equations are correlated. Overall, we believe that these results could be especially relevant for practitioners in cases involving sample selection of unknown form, when the selection process is difficult to model or when exclusion restrictions are not available.

The rest of the paper contains seven sections in addition to this introduction. Section 2 provides a review of the literature. Section 3 presents a general framework and the estimation strategies. Section 4 shows the consistency of many available estimators. The performance of the proposed estimators is tested in Section 5. Here, we present a Monte Carlo study of the finite sample average bias of many relevant cases. In Section 6, we present an empirical application, using the same data of Semykina and Wooldridge (2013, SW) or Lai and Tsai (2016). Finally, Section 7 concludes.

2 Previous literature

The problem of endogenous selection is common in the empirical economic literature using panel data and it has also received attention in theoretical econometrics models. Starting with Verbeek and Nijman (1992), who proposed tests of selection bias either with or without allowance for correlation between the unobserved effects and explanatory variables, a number of proposals considering unobserved heterogeneity and selectivity simultaneously have appeared. Some of them, such as Wooldridge (1995) and Rochina-Barrachina (1999), proposed new methods for estimating the sample selection model with correction under strict exogeneity. Kyriazidou (1997) corrected for selection bias using a semiparametric approach based on a conditional exchangeability assumption and Lai and Tsay (2016) proposed maximum simulated likelihood methods. On the other hand, Vella and Verbeek (1998), Charlier *et al.* (2001) and Semykina and Wooldridge (2010) allowed for endogenous explanatory variables. Finally, Semykina and Wooldridge (2018) proposed estimation procedures for discrete choice panel data models with selectivity.[3]

Dynamics appeared for the first time in Arellano *et al.* (1999), who proposed different solutions for estimating dynamic panel data sample selection models. Next, Kyriazidou (2001) extended her previous proposal to include a lagged dependent variable. 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. Note, however, that the validity of Semykina and Wooldridge's method is based on the validity of the assumption of correlation of the heterogeneity components and the initial condition. Because none of the previous papers suggested a preferred,

[3]In another strand of research, theoretical papers have explored bias-corrected estimators for the static case (Fernández-Val and Vella, 2011).

simplified, or dominant method, our aim here is to provide solutions easily applicable from the point of view of applied practitioners.

Semiparametric alternatives for dynamic panel data sample selection models were studied by Gayle and Viauroux (2007) and Sasaki (2015). Furthermore, maximum likelihood methods were explored by Raymond *et al.* (2007). Note that in the latter case the Montecarlo study is specified with no common covariates between the selection and the outcome equation. A case we'll latter argue that can be estimated with standard uncorrected gmm estimator. Thus with much less assumption and more flexibility.

The various methods have been applied to a number of empirical studies. Charlier *et al.* (2001) studied housing expenditure by households. Jones and Labeaga (2003) selected a sample of non-smokers using the variable addition test of Wooldridge (1995) and then estimated Tobit-type models on the sample of smokers and potential smokers using GMM and minimum distance (MD) methods. González-Chapela (2007) used GMM to estimate the effect of recreational goods on male labour supply. Winder (2004) used instrumental variables to account for endogeneity of some regressors in earnings equations for females. Jiménez-Martín (2006) estimated dynamic wage equations and tested the possibility of differences between strikers and non-strikers. Dustmann and Rochina-Barrachina (2007) estimated females' wage equations extending Rochina-Barrachina (1999). Semykina and Wooldridge (2010, 2013) applied their methods to estimate earnings equations for females. Raymond et al (2010) apply the maximum likelihood methodology develop in Raymond et al (2007) to a model of the occurrence of TPP innovations in Dutch manufacturing enterprises and the extent of these innovations in terms of the share of innovative sales. Knoef and Been (2015) extend Rochina-Barrachina (1999) to an ordinal sample selection models. Interestingly, in the former two cases, the selection and outcome equations have no covariates in common. Note this is one of the cases we argue there is no need to correct selection to obtain consistent estimates. Chang and Trivedi (2015) estimate a model of attrition. Finally, Semykina and Wooldridge (2018) applied discrete choice sample selection panel data models to the analysis of pension coverage among white females in the US.

Because it is likely these approaches will be used more frequently in the future, we believe that it is important to highlight properties, advantages and problems of the various methods, as well as their pitfalls and performance in applied studies. This is precisely what we aim to do in this paper. In more detail, we analyze the necessity for correcting several estimators in static and dynamic models (with either fully exogenous, predetermined or endogenous regressors). We show that corrections are only strictly necessary when both equations have time-varying covariates in common. For example, in the purely AR(1) case, we show the consistency of the AB estimator (and implicitly of the AH estimator) applied to the uncorrected equations and we establish a bound for the system GMM estimator in the worst-case scenario of endogenous selection. Then, we carry out a Monte Carlo exercise to examine the performance of each method under alternative assumptions.

3 A general framework

Consider the following class of panel data models with unobserved heterogeneity:

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

for $i = 1, ..., N$ and $t = 1, ..., T$.

where x is a covariate, that can be either exogenous, predetermined or endogenous, and α_i is an individual heterogeneous component independent of the idiosyncratic error ε_{it} and (assumed for simplicity) independent of x.[4] ρ and β are the parameters of interest.

The combination of different values of ρ and β lead to different models. For example, the assumption $\rho = 0$ leads to a **static panel data model**; when $|\rho| < 1$ and $\beta = 0$ we have a purely stationary **AR(1)**; finally, 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 $|\rho_x| < 1$, $\wp$ is a strictly exogenous covariate, α_i^x is a heterogeneity component and ε_{it}^x is a time-variant error component. 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, z_{it} is a vector of strictly exogenous regressors including a constant and x is a regressor(s), assumed to be independent of z,[5] that may appear also in the outcome equation.[6] We assume that z and x do not have variables in common and so, we can assume z are exclusion restrictions. For convenience we also assume that x has been cleaned of any correlation with η_i.[7] Note that in case $\delta = 0$ the selection equation has no variables in common with the outcome equation. Finally u_{it} is an error term. The observed indicator d_{it} is given by:

[4] Except for the case of the random effects estimator in static models, all the key results of the paper remain unaltered in case we allow x and η_i to be correlated.

[5] The implications of not being independent are the same as the implications of both being in the selection equation.

[6] We also allow the case where x is the lagged outcome y_{t-1}. While this makes identification more difficult, it fits well on our general argument.

[7] For example, we can assume $\eta_i = g(x_i)$ and then add this function as an additional regressor(s).

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

in a way such that $d_{it} = 1$ if y_{it}^* is observed and zero otherwise.[8]

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

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

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

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 the case that 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 two cases: A) the contemporaneous correlation case, when $\vartheta_0 \neq 0$ and $\vartheta_j = 0; 1, 2$; 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 classical or, when necessary, IV methods. In the static case, ($\rho = 0$), with exogenous regressors, any FE and RE estimator is consistent under the maintained assumption that α_i and x are not correlated. Alternatively, when the model is static and x is predetermined or endogenous or the model is dynamic, IV estimators are generally needed. For example, the purely AR(1) model or the dynamic model with covariates are usually estimated by IV, as firstly introduced by Anderson and Hsiao (1982). Arellano and Bond (1991), among others, proposed a more efficient GMM estimator, while Arellano and Bover (1995) extended the previous approach to include equations in levels and proposed the estimation of the whole model using system GMM. Blundell and Bond (1998) extended Arellano and Bover (1995) and noted that in the case of an AR(1) with highly persistent time series correlation, first-differencing could lead to a weak instruments problem (see Roodman, 2009). Then, the use of equations in levels could become important to improve efficiency.

3.1 Estimation under sample selection

3.1.1 The static model case, $\rho = 0$

Estimation in levels: Equation (1) could be estimated in levels by RE. A sufficient condition to consistently estimate the parameters of interest is:

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

[8] Since we will focus on the properties of the estimators for the outcome equation, sometimes we will also exclude η_i from (4).

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: A sufficient condition for OLS to be consistent using differences (or the fixed effect transformation) across time is:

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

Since condition (8) sets no restrictions on how the selection mechanism or the regressors relate to, differencing (1) across time does not only eliminate the problem of correlated individual heterogeneity but also any potential selection problem which operates through. If condition (8) is satisfied, the OLS estimator on the model in time differences provides consistent estimates. Alternatively, if this condition is violated consistent estimation requires considering the selection process.

3.1.2 The AR(1) model

In the small T dynamic case, IV methods are in general necessary.[9] As pointed above, we consider the following estimation options: 2SLS-IV (AH: Anderson & Hsiao, 1982) and, more generally, GMM (AB: Arellano & Bond, 1991; System: Arellano & Bover, 1995; Blundell & Bond, 1998). All of them imply first differencing the data (and combine the estimation using also the equations in levels in the case of the system estimator) and use of internal instruments lagged at least twice, which implies that the selected sample is conditional on observing the outcome for at least three consecutive periods ($d_{it}, d_{it-1}, d_{it-2} = 1$).

Under this condition, for the AH and the AB to be consistent, we need the following orthogonality condition to hold:

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

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

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

which is also stronger than in the general case.

Our initial guess, based on previous work by Arellano *et al.* (1999), is that because the final estimating sample is selected on positives for at least three consecutive previous periods, the need to correct is greatly reduced.[10]

[9] When T is sufficiently large, we can consistently estimate the parameters of the model using the within-groups estimator (see Nickell, 1991).

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

3.2 Estimation under endogenous sample selection

In the presence of endogenous sample selection, researchers are tempted to proceed analogously to the standard static case described by Wooldridge (1995). First, to correct the problem of endogenous selection induced by the correlation of the errors in both equations, and then, to estimate the model. However, as we will show next, there are two distinct cases:

A. When there is some feedback between the (time variant non-deterministic) covariates[11] in the outcome and the selection equation. However, the necessity of 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 A 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 72 in the Appendix A), two (first-differenced static models, see Rochina-Barrachina, 1999) and three (dynamic models, equation 69 in the Appendix A) consecutive observations. We estimate the equation of interest including the appropriate correction(s) using one of the alternative methods described in Table **1**.

  For example, in the case of estimation of the AR(1), 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 (as in 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.[12]

A2. Longitudinal 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 (see Appendix A) estimating bivariate and trivariate probits models of, respectively, the probability that $d_{it} = d_{it-1} = 1$ and $d_{it} = d_{it-1} = d_{it-2} = 1$.

[11] When the common covariates are deterministic or time-invariant there will be no necessity to correct estimates in first-differences and little necessity to correct estimates in levels.

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

- 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 68 in Appendix A). Alternatively, the equation in levels require also two correction terms but, in this case, obtained in a bivariate probit (equation 71 in Appendix A). Note that, since the equations in first differences and levels require different corrections, standard software is not appropriate for obtaining the corrected system estimator, but we suggest to use, for instance, 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.[13] 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. This is so because for the AB

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

and, for the AH

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

Furthermore, the AH and the AB estimators are consistent 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 estimator is not consistent since

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

However, the implied bias is small (especially when the individual heterogeneous components are not correlated) and so, in small samples we are still going to prefer the system estimator. In the model with covariates, the system estimator has a small bias under the same condition, regardless of the nature of the covariates.[14]

[13] 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. However, there are many empirical exercises where these assumptions are not going to be maintained as labor supply models, wage equations, estimation of wage differentials, etc.

[14] Follow-up to result 2: In case we like to correct the bias of the system estimator, we need to correct for selection only the equation in levels. If the correlation between the time-invariant error components is zero and there is no

- 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 not 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 (the latter providing consistent estimates if $cov(x, \alpha_i) = 0$).
- Result 4: When x is not present in the selection equation but is not independent from z is it still possible to avoid bias correction *a la Heckman* by accounting for the relation between x and z, say $E(x|z)$ in the outcome equation.

In Table **1** we summarize all the cases considered and the suggested solutions. We distinguish four static cases and five dynamic models. As we show in the next section, when there are no covariates in common between both equation and they are independent, there is no necessity to correct 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, OTHER
AR(1)	$\|\rho\| < 1$	No	—	nr	No	FD-IV, FD-GMM
Dynamic	$\|\rho\| < 1$	Yes	No	No	No	FD-IV, FD-GMM
Dynamic	$\|\rho\| < 1$	Yes	Yes	No	No	FD-IV, FD-GMM
Dynamic	$\|\rho\| < 1$	Yes	No	Yes	Yes	FD-IV, FD-GMM
Dynamic	$\|\rho\| < 1$	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 account for this correlation.
2. Consistency of the GLS estimator requires the extra assumption of absence of correlation of x with the heterogeneity component in the outcome equation.

feedback between both equations, the bias of the system estimator is small (but not zero). So, when the AB estimator does not work well (small N, large autoregressive coefficient), the system estimator is highly recommended.

4 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 the selection and the outcome equations. For example, for dynamic models the AH and AB estimators are consistent when the outcome and selection equation have no regressors in common, i.e., when all the regressors in the selection equation are exclusion restrictions. The system estimator is an exception and presents a small bias, mainly induced by the correlation between the time-invariant heterogeneous components in the outcome and the selection equations.

4.1 Consistency in the pure autoregressive model

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

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

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

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

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

The exogenous random variables z_{it}, α_i^0, ε_{it}^0, η_i, and u_{it} are assumed to be i.i.d. and independent of each other with finite second moments.[15] We assume that $|\rho| < 1$ and y_{it}^* is the stationary causal solution to the AR(1) model, $y_{it}^* = \frac{\alpha_i}{1-\rho_0} + \sum_{j=0}^{\infty} \rho_0^j \varepsilon_{it-j}$. We also assume that $E(\varepsilon_{it}^0) = E(u_{it}) = 0$. The observed data is the set of y_{it}^* for which $d_{it} = 1$.[16]

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

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

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

[15] We omit x from the selection equation due to its irrelevance for the properties of the estimates in the purely AR(1) case.

[16] We include the lagged latent variable y_{it-1}^* in the right-hand side of the outcome equation, but the reasoning is also valid for the lagged observed variable y_{it-1}.

$$E(s_{ist}y^*_{is}\Delta\varepsilon_{it}(\rho)) = E(s_{ist}y^*_{is}(\Delta y^*_{it} - \rho\Delta y^*_{it-1})) \tag{16}$$
$$= E(s_{ist}y^*_{is}(\rho_0\Delta y^*_{it-1} + \Delta\varepsilon_{it} - \rho\Delta y^*_{it-1})) \tag{17}$$
$$= (\rho_0 - \rho)E(s_{ist}y^*_{is}\Delta y^*_{it-1}) + E(s_{ist}y^*_{is}\Delta\varepsilon_{it}) \tag{18}$$

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

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

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

$$E(d_{it}d_{it-1}d_{it-2}\Delta\varepsilon_{it}|d_{is}, y^*_{is}) = E(d_{it}d_{it-1}d_{it-2}(\Delta\varepsilon^0_{it} + \vartheta_0\Delta u_{it})|d_{is}, y^*_{is}) \tag{20}$$
$$= E(d_{it}d_{it-1}d_{it-2}\vartheta_0(u_{it} - u_{it-1})|d_{is}, y^*_{is}) \tag{21}$$

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

$$E(d_{it}d_{it-1}d_{it-2}\Delta\varepsilon_{it}|d_{is}, y^*_{is}) = \vartheta_0 E(d_{it-2}E(d_{it}d_{it-1}(u_{it} - u_{it-1})|\eta_i, d_{it-2}, d_{is}, y^*_{is})|d_{is}, y^*_{is}) \tag{22}$$

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

$$E\left(d_{it}d_{it-1}(u_{it} - u_{it-1})|\eta_i\right) = E\left(d_{it}d_{it-1}u_{it}|\eta_i\right) - E\left(d_{it}d_{it-1}u_{it-1}|\eta_i\right) \tag{23}$$
$$= E\left(d_{it}u_{it}|\eta_i\right)E\left(d_{it-1}|\eta_i\right) - E\left(d_{it}|\eta_i\right)E\left(d_{it-1}u_{it-1}|\eta_i\right) = 0 \tag{24}$$

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

$$E(s_{ist}y^*_{is}\Delta\varepsilon_{it}(\rho)) = (\rho_0 - \rho)E\left(s_{ist}y^*_{is}\Delta y^*_{it-1}\right) \tag{25}$$

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.

4.1.1 Bound on the bias of the system estimator for the purely AR(1) model

Consider the infeasible level moment conditions $E((y^*_{it}-\rho_0 y^*_{it-1})\Delta y^*_{it-1}) = 0$. The feasible analogue for the moment on the left hand side is $E(d_{it}d_{it-1}d_{it-2}(y^*_{it}-\rho_0 y^*_{it-1})\Delta y^*_{it-1}); d_{it} = d_{it-1} = d_{it-2} = 1$. However, we have verified in Monte Carlo experiments that it is not generally equal to zero in our model. Simulation exercises show that this expectation is, for all reasonable combination of the parameters of the model, very small (see Table A.1 for an illustration) and so is the induced bias.

To be added: Bias of the system estimator in dynamic models

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

4.2.1 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.

$$y^*_{it} = \alpha_i + \rho_0 y^*_{it-1} + \beta_0' x^*_{it} + \varepsilon_{it} \tag{26}$$

$$d_{it} = 1(\eta_i + \gamma_0 z_{it} + u_{it} > 0) \tag{27}$$

$$\alpha_i = \alpha^0_i + \theta_0 \eta_i \tag{28}$$

$$\varepsilon_{it} = \varepsilon^0_{it} + \vartheta_0 u_{it} \tag{29}$$

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.[17] We assume that $|\rho| < 1$ and y^*_{it} is the stationary causal solution to the AR(1) model, $y^*_{it} = \frac{\alpha_i}{1-\rho_0} + \sum_{j=0}^{\infty} \rho_0^j(\beta_0' x^*_{it} + \varepsilon_{it-j})$. We also assume that $E(\varepsilon^0_{it}) = E(u_{it}) = 0$. The observed data is the set of y^*_{it} and x^*_{it} for which $d_{it} = 1$.

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

$$E(s_{ist}y^*_{is}\Delta\varepsilon_{it}(\rho,\beta)) = (\rho_0-\rho)E(s_{ist}y^*_{is}\Delta y^*_{it-1}) + (\beta_0-\beta)'E(s_{ist}y^*_{is}\Delta x^*_{it}) + E(s_{ist}y^*_{is}\Delta\varepsilon_{it}) \tag{30}$$

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

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

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

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

is non-singular.

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

$$
\begin{aligned}
E(s_{ivt}x^*_{iv}\Delta\varepsilon_{it}) &= E(d_{it}d_{it-1}d_{it-2}d_{iv}x^*_{iv}(\Delta\varepsilon^0_{it} + \vartheta_0\Delta u_{it})) && (32)\\
&= E(d_{it}d_{it-1}d_{it-2}d_{iv}x^*_{iv}\vartheta_0\Delta u_{it}) && (33)\\
&= E(d_{it-2}d_{iv}x^*_{iv}\vartheta_0 E(d_{it}d_{it-1}\Delta u_{it}|\eta_i, d_{it-2}, d_{iv}, x^*_{iv})) && (34)\\
&= E(d_{it-2}d_{is}x^*_{iv}\vartheta_0 E(d_{it}d_{it-1}\Delta u_{it}|\eta_i)) && (35)\\
&= 0 && (36)
\end{aligned}
$$

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.

4.2.2 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 $v \leq t-2$ in the above calculations. If we can further assume that x_{iv} is independent of ε_{iv}, u_{iv}, and z_{iv}, then exogeneity will be satisfied with $v = t-1$ as well.

4.2.3 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

$$
\begin{aligned}
E(s_{ist}y^*_{is}\Delta\varepsilon_{it}(\rho,\beta)) &= (\rho_0-\rho)E(s_{ist}y^*_{is}\Delta y^*_{it-1}) + (\beta_0-\beta)'E(s_{ist}y^*_{is}\Delta x^*_{it}) + E(s_{ist}y^*_{is}\Delta\varepsilon_{it}) && (37)\\
E(s_{it}\xi_i\Delta\varepsilon_{it}(\rho,\beta)) &= (\rho_0-\rho)E(s_{it}\xi_i\Delta y^*_{it-1}) + (\beta_0-\beta)'E(s_{it}\xi_i\Delta x^*_{it}) + E(s_{it}\xi_i\Delta\varepsilon_{it}), && (38)
\end{aligned}
$$

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

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

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

4.3 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 follows straightforwardly, but we need to justify it for the RE estimator. The expression for the static model without covariates in common is:

$$y_{it}^* = \alpha_i + \beta x_{it} + \varepsilon_{it} \tag{39}$$

$$d_{it} = 1(\eta_i + \gamma_0 z_{it} + u_{it} > 0) \tag{40}$$

where $x \perp z$. Estimation of the uncorrected RE is carried out in the following selected sample:

$$y_{it}^* = \alpha_i + \beta x_{it} + \varepsilon_{it} \qquad if \qquad d_{it} = 1 \tag{41}$$

where, under endogenous selection, $E(\alpha_i + \varepsilon_{it}|d_{it} = 1) = 0$, provided that $x \perp z$, then 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 (although it affects the standard errors).[18]

4.3.1 Consistency when $x \not\perp z$

As an extention of the case above, we consider the case in which x is not present in the selection equation but $x \not\perp z$. We'll show that the uncorrected estimators are still consistent provided we control for the relation between x and, say, z', the covariates in z that have some relationship with x. So, let us consider the following control function approach in the spirit similar to Olsen's (1980) solution for sample selection in static models.

- Consider $z' \in z$ such that $cov(x, z') \neq 0$. Then, under very standard assumptions, adding $E(x|z')$ [or more generally $E(x|z)$] to the outcome equation corrects the bias of the parameter of x.

  So, for the case of the static model estimated in levels described , we adjust equation 41:

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

  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.

[18] Note that this result also applies to cross-sectional analyses.

Finally, note this result can be applied to all models, either static or dynamic, in which the covariates in both equations are distinct but not independent.

4.4 Consistency in panel data sample selection models when $\delta \neq 0$

When at least a 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 Wooldrigde (1995), bias correction induced by endogenous sample selection implies adding univariate corrections if the sample is conditional on only one observation (random effects strategy in the static model). In Appendix B we show that Wooldrige'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, as 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).

4.4.1 The correction procedure

We describe the correction procedures in two steps:

- **Step 1. Estimation of the corrections**

  (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 [19]. 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 static models, or trivariate probit models to correct dynamic models.[20] See Appendix B for details.

  Important Results: A follow up from cases where there is no need to correct is the fact that omission of any regressor in the selection equation, $z' \in z$, such that $z' \perp x$, does not affect the consistency of the corrected estimates. **An immediate implication of this result is that the selection equation can be misspecified in some cases.**

[19] recall we do not allow correlation between x and the heterogeneity component of the selection equation. Otherwise we will follow Mundlak's approach to correct the problem

[20] The order of the appropriate correction increases accordingly in AR(p) models.

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

  (i) **Errors contemporaneously correlated only.** In this case, all the estimators considered in this paper (FE, FD, RE, for the static model, and AH, AB, system for the dynamic model) require corrections derived after adjusting univariate year-by-year probits.

  In the RE and level equations of the system estimator the corrections are introduced in levels. In first-differenced models, the corrections are introduced in first-differences. Finally, in the FE estimator, the correction is introduced in within-differences.

  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):

$$y_{it} = \rho y_{it-1} + x_{it}\beta + \sigma\lambda\left(H_{it}\right) + e_{it} \tag{42}$$

$$\Delta y_{it} = \rho \Delta y_{it-1} + \Delta x_{it} + \bar{\sigma}(\lambda\left(H_{it}\right) - \lambda\left(H_{it-1}\right)) + \Delta e_{it} \tag{43}$$

  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.** In this case, the number of periods an observation is conditional on is critical in determining the appropriate correction. As described in Appendix B, in static models estimated by GLS we only need to add a single correction; in static models estimated by FD we need to add two correction terms obtained 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 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 (for examples, xtabond2).

  As a matter of example, we show the corrections needed for the system estimation (see Appendix B for a description of the notation). [21]

$$y_{it} = \rho y_{it-1} + x_{it}\beta + \bar{w}_i\psi + \sigma_0\lambda\left(H_{it}, H_{it-1}, \varrho_{t,t-1}\right) + \sigma_{-1}\lambda\left(H_{it-1}, H_{it}, \varrho_{t,t-1}\right) + e_{it} \tag{44}$$

[21] Note that when x is endogenous the corrections need to be instrumented using the same lag order used to instrument the covariate.

$$
\begin{aligned}
&\Delta y_{it} = \rho \Delta y_{it-1} + \Delta x_{it} + \bar{\sigma}(\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) \\
&-\lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right)) \\
&+\bar{\sigma}_{-2}\lambda\left(H_{it-2}, H_{it-1}, H_{it}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) + \Delta e_{it}
\end{aligned} \tag{45}
$$

where $\varrho_{t,t-s}$ denotes the correlation between errors in period t and $t-s$ and the function 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.[22] Finally, a standard t-test of significance of the correction term (or a Wald test, in case of multiple lambda's) stands for an approximate test of endogenous selection (Wooldridge, 1995).

4.4.2 Construction of the corrections

Univariate 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 $\hat{\lambda}_{it}(z_{it}\hat{\gamma})$. When we allow correlation between z_{it} and η_i, we can rely on Mundlak (1978) and assume, for instance, $\eta_i = \tilde{z}_i\varphi$, where $\tilde{z}_i$ is the vector of individual means of z_{it}, and we, again, can estimate a probit for each period and compute $\tilde{\lambda}_{it}(z_{it}\tilde{\gamma} + \tilde{z}_i\tilde{\varphi})$, which is then introduced in a second step as before.[23]

Bivariate or trivariate corrections Assuming the errors $\nu_{it}, \nu_{it-1}, \nu_{it-2}$ are jointly normal we can estimate bivariate or trivariate probits in order to construct the bivariate and the trivariate correction. See the Appendix B for the details.

More general corrections In Appendix B we describe semiparametric estimates of the correction than can overcome the failure of, for instance, the normality assumption.

5 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:

[22] See the Appendix C.

[23] 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 $\eta_i = \pi_1 z_{i1} + \pi_2 z_{i2} + ... + \pi_T z_{iT}$ 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 (the $z's$) and we predict the index $\hat{d}^*_{it}$. Then, in a second stage, we estimate the structural parameters by within-groups, MD or GMM and compute the correction terms based on these two-stage 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.

$$d_{it}^* = a - z_{it} - \delta x_{it} - \eta_i - u_{it} \tag{46}$$

$$d_{it} = 1[d_{it}^* > 0] \tag{47}$$

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

$$y_{it}^* = (2 + \alpha_i + \varepsilon_{it})/(1 - \rho) \text{ if } t = 1 \tag{48}$$

$$y_{it}^* = 2 + \rho y_{it-1}^* + \beta x_{it} + \alpha_i + \varepsilon_{it} \text{ if } t = 2, ..., T \tag{49}$$

$$y_{it} = y_{it}^* \text{ if } d_{it} = 1 \tag{50}$$

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 minimise any problem with initial conditions.[24] We consider the following process for x:

$$x_{it} = (0.5 + \wp_{it} + \alpha_i^x + \varepsilon_{it}^x + \kappa(\alpha_i + \varepsilon_{it}))/2 \quad if\, t = 1 \tag{51}$$

$$x_{it} = 0.5 + 0.5x_{it} + \wp_{it} + \alpha_i^x + \varepsilon_{it}^x + \kappa(\alpha_i + \varepsilon_{it}) \quad if\, t > 1 \tag{52}$$

and we let κ varying so that: (i) when $\kappa = 0$, x is fully exogenous; alternatively, (ii), alternative whe $\kappa = 0.5$, x is either endogenous or predetermined (in which case ε_{it} is replaced by ε_{it-1})

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

$$\wp_{it} \sim N(0, \sigma_\wp) \text{ with } \sigma_\wp = 1 \tag{53}$$

$$z_{it} \sim N(0, \sigma_z) \text{ with } \sigma_z = 1 \tag{54}$$

$$\eta_i \sim N(0, \sigma_\eta) \text{ with } \sigma_\eta = 1 \tag{55}$$

$$u_{it} \sim N(0, \sigma_u) \text{ with } \sigma_u = 1 \tag{56}$$

$$\alpha_i = \alpha_i^0 + 0.5\eta_i, \alpha_i^0 \sim N(0, \sigma_{\alpha^0}) \text{ with } \sigma_{\alpha^0} = 1 \tag{57}$$

$$\varepsilon_{it} = \varepsilon_{it}^0 + \vartheta_0 u_{it} + \vartheta_1 u_{it-1} + \vartheta_2 u_{it-2}, \varepsilon_{it}^0 \sim N(0, \sigma_{\varepsilon^0}) \text{ with } \sigma_{\varepsilon^0} = 1 \tag{58}$$

$$\alpha_i^x \sim N(0, \sigma_{\alpha^x}) \text{ with } \sigma_{\alpha^x} = 1 \tag{59}$$

$$\varepsilon_{it}^x \sim N(0, \sigma_{\varepsilon^x}) \text{ with } \sigma_{\varepsilon^x} = 1 \tag{60}$$

Where, in the case A1 of contemporaneous correlation, we set $\vartheta_0 = 0.5; \vartheta_1 = \vartheta_2 = 0$. These

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

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 os serially cross-correlated errors we set $\vartheta = 0.5; \vartheta_1 = 0.5/2; \vartheta_2 = -0.5/3$.

5.1 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 the lost is less 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 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) Static model with an exogenous x

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

(v) Dynamic model with an endogenous covariate not present in the selection equation: $\rho = 0.25; \rho = 0.75$, $\beta = 1$ $\delta = 0$

(vi) Dynamic model with an endogenous covariate also present in the selection equation: $\rho = 0.25; \rho = 0.75$, $\beta = 1 = \delta$

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. Selection of the instruments is a crucial step of our simulation study. In both cases we select the instruments as follows: we use lags from $t-2$ backwards for first-differenced equations, although we also evaluate the performance of the estimates with a restricted set of instruments. We use the lagged first difference of the outcome as an additional instrument for the equation in levels 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.[25]

[25] We also use Roodman's proposal to collapse the number of instruments and we get very similar results, available upon request, in the empirical applications.

Table 2: Average bias and RMSE in the static model. T=7; 500 replications

x in selection	Corrected	FE estimator av. bias	FE estimator RMSE	FE estimator ERF correction	FD estimator av. bias	FD estimator RMSE	FD estimator ERF correction	RE (GLS) estimator av. bias	RE (GLS) estimator RMSE	RE (GLS) estimator ERF sel. term
Panel A: N = 500; endogenous selection, $cov(x,z)=0$										
No	No	.00023	.02143		-.00063	.03008		.00057	.01510	
Yes	No	-.04741	.05365		-.04269	.05542		-.06408	.06702	
Yes	Yes[1]	.00098	.02948	.95	.00079	.03890	.77	-.00344	.02447	1
Panel A: N = 500; exogenous selection, $cov(x,z)=0$										
Yes	Yes[1]	.00104	.02611	.05	.00076	.03484	.05	.00096	.02181	.07
Panel B: N = 5000; endogenous selection, $cov(x,z)=0$										
No	No	-.00008	.00700		.00025	.00933		.00010	.00518	
Yes	No	-.04694	.04763		-.04124	.04264		-.06397	.06426	
Yes	Yes[1]	.00202	.00966	1	.00225	.01279	1	-.00235	.00812	1
Panel B: N = 5000; exogenous selection, $cov(x,z)=0$										
Yes	Yes[1]	.00013	.00840	.04	.00054	.01126	.06	.00009	.00669	.04
Panel C: N = 500; endogenous selection, $cov(x,z)\neq 0$										
No	No	-.02532	.03046		-.03147	.03993		-.02107	.02509	
No	Yes[2]	-.00107	.02044	.50	-.00178	.02896	.38	-.00038	.01540	.74
Panel C: N = 5000; endogenous selection, $cov(x,z)\neq 0$										
No	No	-.02533	.02585		-.03146	.03235		-.02122	.02164	
No	Yes[2]	-.00035	.00635	1	.00000	.00901	1	-.00022	.00484	1

1. In Panels A y B the correction is obtained from a year by year probit with z as a covariate.
2. In Panel C the correction is $E(x|z)$, being $cov(x,z')\neq 0$ where $z'\in z$.

5.2 Simulation results for static models

In this section we present simulations for static models, all of them under the assumption that the errors in both equations are contemporaneously correlated.

5.2.1 Static model with $\delta = 0$ and $x \perp z$

This case corresponds to the basic static model with a covariate x, absent from the selection equation and unrelated to z, i.e., $cov(x,z)=0$. We consider three estimators: FE, FD and RE estimated by GLS. The results are reported in the first row of Panels A and B in Table **2**. They show that the average bias is almost zero, regardless of the initial sample size, small (Panel A) or large (Panel B). According to the RMSE criterion, since in our experiment $cov(x,\alpha_i)=0$, the RE is our preferred method. Otherwise, the FE estimator will be mildly preferred to the FD one.

5.2.2 Static model with $\delta \neq 0$ and $x \perp z$

The next simulations correspond to the basic static model with a covariate x included in both equations or static model with observed feedback. We again consider three estimators, FE, FD and RE estimated by GLS, and we present two sets of estimates, uncorrected (to make evident the

selection bias of the uncorrected estimates) and corrected for selection. The results are reported in the second and third rows of Panels A and B in Table **2**.

When we do not correct for sample selection bias (second row of Panels A and B of Table **2**), all three estimators are biased regardless of the sample size. The results of row 3 of Panels A and B show the effects that correction *a la Wooldridge* has when at least a covariate is included in the selection and outcome equations: average bias is very small regardless of the sample size.[26] As in the previous subsection, since $cov(x, \alpha_i) = 0$, the RE presents a lower RMSE, i.e., is, as expected, more efficient than either the FE or the FD estimator.

Sample selection test. We also 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 initial 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.

5.2.3 Static model with $\delta = 0$ and $x \not\perp z$

A very interesting case arises when x is not included in the selection equation but it is correlated with some variables includes in the vector z, say z'. We show in Panel C of Table **2** that the uncorrected estimates are biased. In these circumstances, one would be tempted to follow standard sample selection approach, and add a Heckman's type correction to the outcome equation. However, as we have described in section 4.3.1 this is not strictly necessary. In order to control the bias we add to the outcome equation an estimate of $E(x|z')$

The simulated results for this procedure are presented in Panel C of Table **2**. Our proposal takes out practically all the bias regardless of the sample size, but especially when the sample is large, $N = 5000$ in our case. The specification test shows some lack of size when $N = 500$, although this problem disappears as the sample grows.

5.3 Simulation results for AR(1) model

5.3.1 Basic results

Table **3** presents results for the AR(1) model for three values of the autoregressive parameter: 0.25, 0.50 and 0.75 under the assumption that the errors are only contemporaneously correlated.[27] We

[26] We obtain the same qualitative results when the lagged outcome is included in the selection equation and the outcome equation is dynamic in nature.

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

report results for both the AB and the system estimators constructed under competing assumptions about the selection process: (a) non-endogenous selection; (b) endogenous selection without correction. The initial degree of sample selection is 15 per cent, while the fraction of the sample lost is much larger (around 1/3 of the observations on average).

Let us start reviewing the results without endogenous selection, reported in columns (1) and (2). When the initial sample (before selecting the observations) is small ($N = 500$) the bias of the AB grows with the autoregressive parameter (for both selection models, A and B) and becomes sizable when $\rho = 0.75$.[28] As we increase the sample size ($N = 5000$), the average bias of the AB estimator is reduced substantially and only remains noticeable for $\rho = 0.75$. Alternatively, the system estimator, which is also consistent in this case, shows a very small bias for $N = 500$ (never exceeding one per cent), even smaller when $N = 5000$. Figure **1** confirms these results with a sample size varying from $N = 200$ to $N = 5000$ in the absence of any sort of selection (estimators labeled AB all and system all).

When endogenous sample selection is considered (see columns (3) and (4) for, respectively, the uncorrected AB and system estimators), we do not detect any significant change in the bias results 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 estimator always shows a very small bias (between 1 per cent for $\rho = 0.25$ and 2.25 per cent for $\rho = 0.75$). Note that the bias becomes more evident as the sample size grows (see Figure **1**). As a sort of compensation, the standard errors for the system estimator always tend to be substantially smaller.

Some additional conclusions can be drawn when varying the sample size (Figure **1**). When $N = 200$, the AB estimator shows sizable bias, which decreases as N increases. The system 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 estimator being smaller Below this threshold, and larger above it. Therefore, we may conclude that for moderate and small samples (say, below the range 1000-1500), the system estimator is highly recommended because of the likely smaller bias as well as smaller variance.

Finally, as shown in Figure **2**, when α_i and η_i are not correlated, the bias of the system 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 (see Table A.1 for an illustration).[29] In the next section, we describe a control function approach to account for the correlation between α_i and η_i, thereby reducing to a minimum the bias

[28] 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.

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

Table 3: Average bias and RMSE in the purely AR(1) model. T=7; 500 replications

	Estimates with the full sample				Estimates with the selected sample						
	AB estimator		system		AB estimator		system		system, level eq. corrected[1]		
ρ	av. bias	RMSE	av. bias	RMSE	av. bias	RMSE	av. bias	RMSE	av. bias	RMSE	ERF[2]
Panel A: N=500											
0.25	–.00573	.04067	.00049	.03196	-.01407	.05572	-.00339	.04362	-.00038	.00183	.97
0.50	–.01173	.05603	.00205	.03680	-.03172	.08264	-.00825	.05146	-.00147	.00244	.97
0.75	–.04250	.10140	.00836	.04455	-.10030	.16542	-.00909	.06521	-.00081	.00375	.87
Panel B: N=5000											
0.25	-.00132	.01189	-.00032	.00945	-.00113	.01708	-.00366	.01293	-.00022	.00015	1
0.50	-.00196	.01635	-.00007	.01131	-.00249	.02455	-.00952	.01735	-.00180	.00020	1
0.75	-.00451	.02900	.00071	.01366	-.00856	.04317	-.01821	.02612	-.00697	.00036	1

1. Corrected System estimator. Control function approach to correct the level equations only.
2. ERF of the correction term. The correction term has been obtained from a fixed effect first stage regression.

of the system estimator.

5.3.2 A simple procedure for bias reduction of the uncorrected system estimator in AR(1) model or dynamic models with $\delta = 0$ and $x \perp z$

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

- Step 1: Provided the selection equation has an exclusion restriction, obtain a consistent estimate of the fixed effects ($\hat{\eta}_i$) in the selection equation using a linear probability model.[30]
- Step 2: Add $\hat{\eta}_i$ to the following equation in levels to control the correlation between the time-invariant errors.

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

  where we can assume that α_i^* is not correlated with the time invariant component in the selection equation.

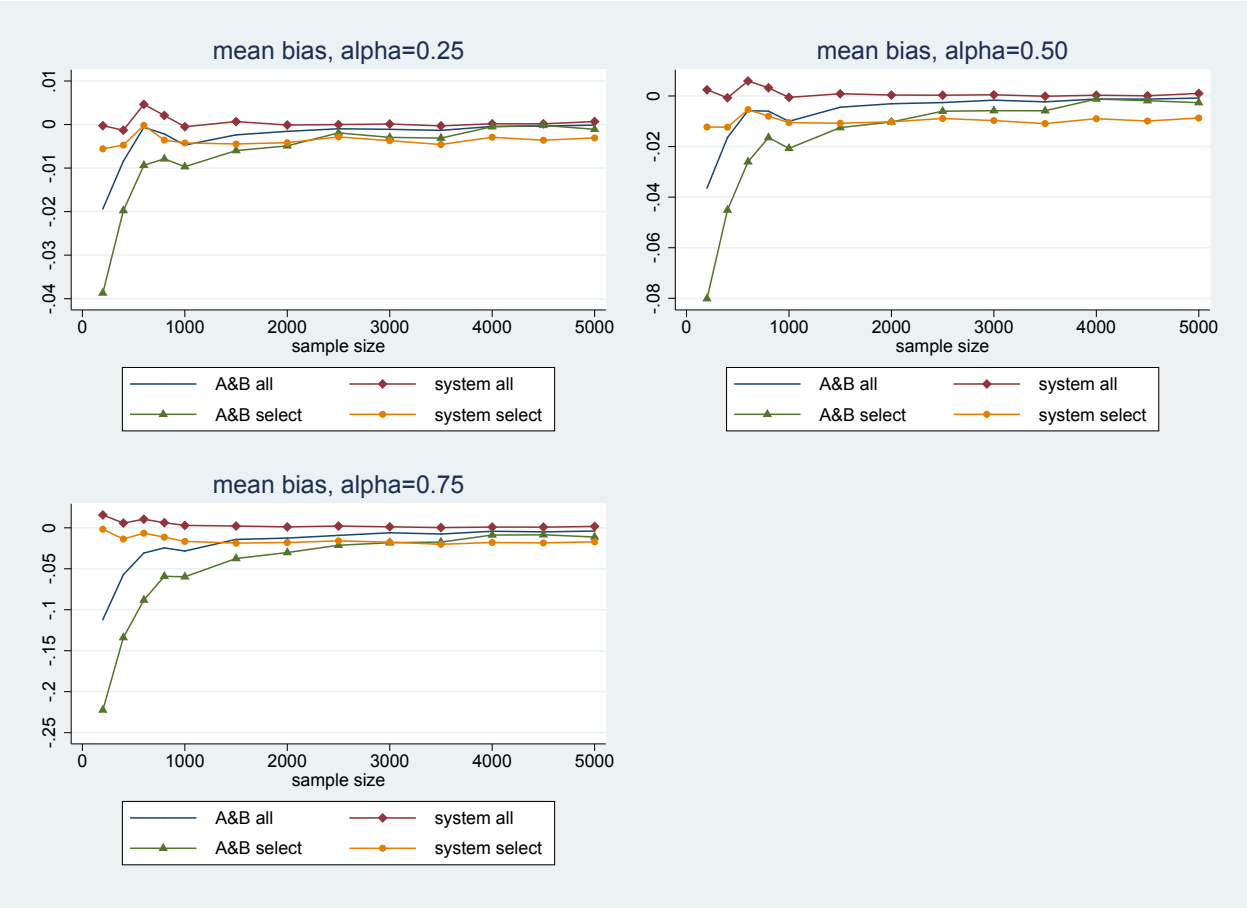
[30] Olsen (1980) proposed a similar method with a least squares correction in a cross-section context.

- Step 3: Estimate the previous equation combining the uncorrected equations in first differences and the corrected equations in levels.

A simple t-test of the null $\theta = 0$ stands for a test of endogenous selection. As in the previous case, corrected standard errors can be computed using the delta method or bootstrapping. If we cannot reject the null hypothesis, the individual heterogeneous components are uncorrelated, so the only potential source of endogenous selection is the correlation of the time-variant errors. Therefore, the only remaining problem for the consistency of the system GMM estimator is the potential correlation between the time varying errors of both equations. However, we show in Table A.1 that this correlation does not generate much bias.

We present in the last three columns of Table **3** simulations of average bias, RMSE and ERF for the test. With respect to the uncorrected system estimators, the magnitude of the average bias (and the RMSE) is reduced between 1/2 and 2/3, depending on the autocorrelation coefficient and the sample size. The test of correlation between the heterogeneity components has strong size, except when the sample size is small ($N = 500$) and $\rho = 0.75$.

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



Notes.

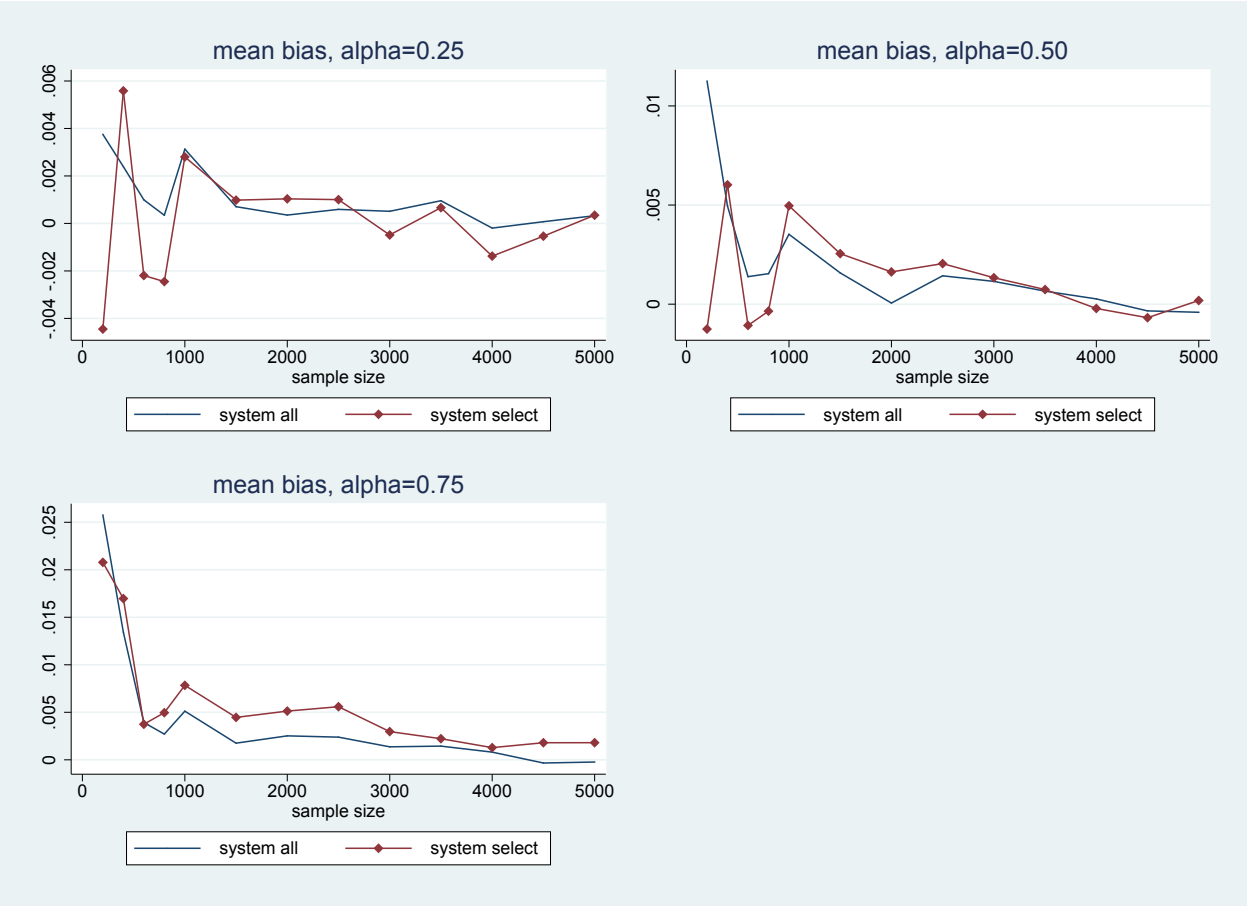
AB all: AB GMM estimates using the full (NxT) sample (no selection process).

system all: System GMM estimates using the full (NxT) 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.

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



Notes.
system all: System GMM estimates with the full sample (no-selection).
system select: Uncorrected system GMM estimates with the selected sample under endogenous selection due to correlation of the time-varying errors.

5.3.3 Additional results for the AR(1)

We have done several Monte Carlo exercises with cases departing from the basic assumptions of the purely AR(1) model.[31] We consider the following cases: (a) varying the longitudinal dimension of the panel; (b) increasing the percentage of selection (from 0.15 to 0.25); (c) increasing the ratio of the variances to $\frac{\sigma_\alpha^2}{\rho_\varepsilon^2} = 2$; (d) reducing the correlation between the errors (the correlation parameter is reduced from 0.5 to 0.25); (e) and, finally, non-stationary time varying errors and correlation of the time-varying error components. In particular, we allow the variance of the time-varying errors in (1) and (2) to vary over time[32] and we also allow the correlation coefficient between the

[31] These results are not reported in the paper, but they are available from the authors on request.
[32] We multiply either ε_{it} or u_{it} by a time-varying Bernoulli process taking either 1 or 2.

time-varying errors in (1) and (2) to vary over time.[33] All these sensitivity exercises confirms 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 while the system estimator does not. All these results imply that the system estimator is specially recommended when the sample size is small or even moderate (below 1000 and 1500 individuals) and less necessary when the sample size is very large.

5.4 The dynamic model with covariates

This section is devoted to Monte Carlo exercises for dynamic models with a covariate that can either be or not included in the selection equation. This variable can be either exogenous, predetermined or endogenous. We present two simulation exercises of dynamic models with an endogenous covariate and another one with an exogenous variable.

5.4.1 The dynamic model with an endogenous covariate not included in the selection equation and $cov(\varepsilon_{it}, u_{is} = 0; s < t)$

We present simulations of a dynamic model with an endogenous covariate x, but independent of z (in case they are not independent we will follow the procedure described in section 4.3.1), the covariate in the selection equation. The key results obtained are shown in the first two columns (for $\rho = 0.25$ and $\rho = 0.75$) of Panels A and B in Table **4**. They find that the AB is consistent when there is an endogenous covariate in the outcome equation not present in the selection equation. The small biases found with $N = 500$ decrease as the sample size increases (they practically disappear when $N = 5000$). The system estimator, although not consistent, has a very small bias regardless of the sample size. More importantly, the RMSE is smaller than the AB case, even when the sample is large ($N = 5000$). Note however, that for very large samples the latter remark is no longer true, since the bias of the AB estimator goes to zero while the bias of the system one does not.[34]

5.4.2 The dynamic model with an endogenous covariate included in the selection equation and $cov(\varepsilon_{it}, u_{is} = 0; s < t)$

Now, we focus on a dynamic model with a covariate x present in both, the outcome and the selection equations. We present both uncorrected and corrected estimates. The uncorrected results are reported in the third and four rows and the corrected ones in fifth and six rows of Panels A and B in Table **4**. The uncorrected estimates are biased regardless of the sample size, which shows the necessity of correcting for sample selection when there is at least a common covariate in both equations. The necessity to use GMM implies correcting the outcome equation with current, lagged and lagged twice lambda correction terms. Furthermore, since x is endogenous, these additional

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

[34] All these results also apply to the case where x is predetermined or exogenous. We do not report them, but they are available on request for interested readers.

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

			AB					SYSTEM				
x in	Corrected	value	ρ		β		λ test	ρ		β		λ test
selection		ρ	av. bias	RMSE	av. bias	RMSE	ERF	av. bias	RMSE	av. bias	RMSE	ERF
					Panel A: N=500; endogenous selection							
No	No	.25	-.01089	.03226	.00984	.03930		-.00600	.02390	.00744	.03268	
No	No	.75	-.02905	.05914	-.00869	.05054		-.00507	.02241	.00427	.03049	
Yes	No	.25	-.02356	.03762	-.03885	.05751		-.01021	.02816	-.03677	.05508	
Yes	No	.75	-.03337	.04648	-.05507	.0714		-.01551	.03571	-.04409	.06096	
Yes	Yes[1]	.25	-.03444	.04512	.02349	.05467	.51	-.01842	.03210	.01293	.04797	.42
Yes	Yes[1]	.75	-.03866	.05054	-.00031	.05285	.45	-.01517	.03312	.00440	.04587	.43
					Panel A: N=500; exogenous selection							
Yes	Yes[1]	.25	-.01647	.03833	.00071	.04697	.04	.00034	.03087	.00088	.03877	.04
Yes	Yes[1]	.75	-.02452	.04615	-.01315	.05389	.04	.01352	.03169	.00225	.03915	.05
					Panel B: N=5000; endogenous selection							
No	No	.25	-.00107	.01007	.00071	.01089		-.00195	.00780	.00067	.00930	
No	No	.75	-.00405	.01647	-.00198	.01481		-.00468	.00843	-.00143	.00907	
Yes	No	.25	-.01277	.01576	-.05529	.05678		-.00525	.00999	-.04928	.05075	
Yes	No	.75	-.02173	.02389	-.06365	.06512		-.01788	.02067	-.05621	.05772	
Yes	Yes[1]	.25	-.01922	.02124	-.00757	.01683	1	-.01135	.01401	-.0048	.01442	1
Yes	Yes[1]	.75	-.02332	.02528	-.01912	.02477	1	-.01478	.01773	-.00998	.01711	1
					Panel B: N=5000; exogenous selection							
Yes	Yes[1]	.25	-.00200	.01076	-.00047	.01392	.06	-.00028	.00879	-.00076	.01239	.05
Yes	Yes[1]	.75	-.00282	.01171	-.00231	.01561	.06	.00184	.01090	-.00042	.01290	.04

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

terms need to be instrumented using lag two and backward lags. The effects of sample correction on the magnitude of the bias reduction is very important, specially in the case of β. These reductions as well as decreases of the RMSE are related to the sample size.

Sample selection test When the sample is small the ERF is small (around 0.50), so the sample selection test fails to detect the presence of endogenous sample selection for both estimators. As the sample increases the performance of the test improves substantially with an ERF close to 1. When the null is true the ERF fall in a range of 0.38 (lowest) to 0.06 (highest).

5.4.3 The dynamic model with an exogenous covariate not included in the selection equation and $cov(\varepsilon_{it}, u_{is} \neq 0 s < t)$

In this section we explore the estimation of a sample selection model when the time-variant errors are cross-serially correlated $cov(\varepsilon_{it}, u_{is} \neq 0; s \leq t)$. As shown we the two equation do not have covariates in common and independent, there is no necessity to correct the estimates, even we 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 **5**. Although this is a very difficult case the results show very small bias of the uncorrected estimator in both samples.

5.4.4 The dynamic model with an exogenous covariate included in the selection equation and $cov(\varepsilon_{it}, u_{is} \neq 0; s \leq t)$

Another feature of these models that we like to explore is the presence of time-variant errors are cross-serial correlation between the time-variant errors of both equations, i.e., $cov(\varepsilon_{it}, u_{is} \neq 0; s \leq t)$. We show in Appendix B that when there are common covariates the estimation of the model either by GMM of system GMM requires multiple correction terms. In the particular case of the system GMM, we have to add two correction terms obtained in trivariate probit models to the first-differenced equations and two additional terms obtained in bivariate probit models to the equation in levels. Moreover, we have to use a Wald test instead of a typical t-test to check for sample selectivity.

The simulation results corresponding to these exercises are reported in Table **5**. They 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 decreases as N grows. However, the bias of the corrected estimator is very small and decreases with N. 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 there is not correlation the ERF stabilizes around 0.06 both for the AB and system estimators.

Table 5: Average bias and RMSE 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	Corrected	value	ρ		β		λ's test	ρ		β		λ's test
selection		ρ	av. bias	RMSE	av. bias	RMSE	ERF	av. bias	RMSE	av. bias	RMSE	ERF
					Panel A: N=500; endogenous selection							
No	No	.25	-.00040	.02253	-.00323	.02680		.01737	.02736	.00682	.02693	
No	No	.75	-.01324	.02571	-.00515	.02765		.00783	.01978	.01083	.02764	
Yes	No	.25	-.01926	.03446	-.04024	.05064		.00233	.02769	-.03157	.04421	
Yes	No	.75	-.02664	.03769	-.04468	.05447		-.00886	.02574	-.03542	.04751	
Yes	Yes	.25	-.00845	.03196	-.00150	.04544	.48	.02073	.03419	.00606	.04422	.76
Yes	Yes	.75	-.01972	.03584	-.00844	.04478	.48	-.00297	.02120	.00636	.04243	.70
					Panel A: N=500; exogenous selection							
Yes	Yes	.25	-.01460	.03212	-.00326	.04122	.17	-.00109	.02466	-.00031	.03796	.29
Yes	Yes	.75	-.01431	.02843	-.00583	.04164	.16	.00337	.01904	.00191	.03819	.29
					Panel B: N=5000; endogenous selection							
No	No	.25	.00608	.00943	-.00153	.00749		.01688	.01817	.00649	.00964	
No	No	.75	-.00498	.00857	-.00190	.00769		.00394	.00729	.00951	.01182	
Yes	No	.25	-.00975	.01291	-.03855	.03954		.00243	.00844	-.03249	.03363	
Yes	No	.75	-.01513	.01709	-.04010	.04111		-.01583	.01728	-.03859	.03963	
Yes	Yes	.25	.00923	.01330	.00153	.01456	1	.02330	.02470	.00465	.01404	1
Yes	Yes	.75	-.00123	.00841	-.00250	.01409	1	-.00420	.00755	.00041	.01225	1
					Panel B: N=5000; exogenous selection							
Yes	Yes	.25	-.00087	.00788	.00236	.01308	.06	.00091	.00746	.00218	.01193	.05
Yes	Yes	.75	-.00052	.00753	.00110	.01332	.07	.00100	.00577	.00111	.01203	.06

Testing univariate corrections vs multiple corrections

			AB					SYSTEM				
x in	Corrected	value	ρ		β		$xtra\lambda$'s test	ρ		β		$xtra\lambda$'s test
selection		ρ	av. bias	RMSE	av. bias	RMSE	ERF	av. bias	RMSE	av. bias	RMSE	ERF
					Panel C1: N=500; endogenous selection but $cov(\varepsilon_{it}, u_{is} = 0; s < t)$							
Yes	Yes	.25	-.01951	.05030	-.00518	.0408	.13	-.00139	.03862	-.00405	.03976	.20
Yes	Yes	.75	-.03438	.06223	-.01231	.04536	.14	.00264	.03034	-.00140	.04139	.22
					Panel C2: N=5000; endogenous selection but $cov(\varepsilon_{it}, u_{is} = 0; s < t)$							
Yes	Yes	.25	-.00196	.01439	-.00118	.01409	.08	-.00245	.01127	-.00081	.01268	.20
Yes	Yes	.75	-.00465	.01773	-.00156	.01508	.06	.00058	.00967	-.00006	.01279	.20

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.

5.4.5 Testing univariate vs multivariate corrections

Our final Monte Carlo exercise compares univariate tests of selection bias presented in Panels A and B of Table **5** with multivariate ones. In 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 in the equation, respectively). In these circumstances, it is easy to show that the coefficient of the lagged twice trivariate lambda in the first-differenced equations and the coefficient of the lagged bivariate lambda in the equation in levels should be equal to zero. Then, a simple t-test in the corrected AB estimator or a Wald test in the corrected system estimator stand for checks of longitudinal correlation between the errors in the outcome and the selection equations. We obtain the expected results as reported in Panel C of Table **5**.

6 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 SW. The second uses consumption data from the Spanish Continuous Family Expenditure Survey (ECPF from now on) to adjust myopic and rational addiction 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, and Becker *et al.*, 1994). Our objective here is twofold. First, we estimate a myopic model of consumption trying to mimic our autoregressive proposals. Second, we adjust a rational addiction model to compare to Jones and Labeaga (2003).

6.1 Estimating female earnings equations

In this first application, we employ the same data used in SW, which were also used by Lai and Tsai (2016).[35] 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 Semykina and Wooldridge, 2013). The results for the pure autoregressive model are presented in Table **6**. Then, we extend the model in Table **7** to include age, aage squared and number of years of education. The first column in Table **6** presents first-differenced IV estimates. Alternatively, column (1) in Table **7** reports the SW estimator. Columns (2) and (3) in both tables report AB and system results obtained in the selected sample, but when we do not correct the earnings equation. Column (4) in both tables adds a correction for

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

the correlation between the unobserved heterogeneous components. In column (5) of Table **6** we present a year-by-year correction only for the equation in levels. Alternatively, in columns (5) and (6) of Table **7** we present year-by-year probit corrections under the assumption that the errors in both equations are contemporaneously correlated. [36]

In Table **7**, we consider the demographic variables to be strictly exogenous and we instrument the lagged log of the dependent variable using all available instruments for both the equations in levels and first-differences. The number of overidentifying restriction is 65 in 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 (see Roodman, 2009).[37]

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.[38] Adding a correction for the correlation of the unobserved heterogeneity components (see column 4), barely changes the coefficient. Alternatively, adding a year-by-year correction in either the equation in levels or in all equations mildly increases the autoregressive parameter. Note, however, that the selection terms are found to be jointly significant.

The autoregressive coefficient (as well as its standard error) remains practically identical in the extended model in Table **7** compared to the pure autoregressive case, and it is substantially lower than the one obtained by SW. Given that the all first stage variables are either time-invariant (education) or deterministic (age and age square) the uncorrected first differences estimates are consistent. So, the result in column (5) are not necessary (if we assume that the first stage regression is correct). The proposed corrections of the system 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 that paper). Furthermore, another dynamic earnings model using the PSID for the 1968-81 period, in this case for males (Holtz-Eakin *et al.*, 1988), yields a similar result for the coefficient of lagged log earnings.

It is also important to note that our age and education estimates are very different from the results in SW, but they are in line with those found in the previous literature using similar data.

[36] All the AB and system GMM estimates, except those reported in column (5) of Table **6**, were obtained using the stata xtabond2 package (see Rodman, 2006). The estimates reported in column (5) have been obtained using a modified version of xtabond2 that only includes the correction in the equation in levels. Note, however, that these estimates can be also obtained using the Stata *gmm* routine.

[37] For instance, when we use up to the fourth lag instead of all lags of the log hourly earnings, we obtain the following coefficients: 0.178, 0.093, 0.020 and -0.0002 for the lagged dependent variable, education, age and age squared, respectively. They compare with those in column 3 of Table **7**.

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

The coefficients of age, age squared and education have the expected signs, with a quadratic profile of age showing increasing earnings at a decreasing rate. The return to education we get is more in line with the average return to education for females for the US usually found in the literature (see Card, 1999, Harmon *et al.*, 2003 or Polachek, 2008). Regarding endogenous selection, we do not detect endogenous selection due to correlation between the time-invariant heterogeneity components (column (5) and (6) in Table **7**).

Table 6: **AR(1) log hourly earnings equation**

	(1)	*(2)*	*(3)*	*(4)*	*(5)*
	2SLS-IV	*No correction*	*No correction*	*Het. components correction of lev eq. only*	*yby correction of lev eq. only*
		AB	*system*	*system*	*system*
Lag log hourly earnings	0.1522**	0.1029**	0.1798***	0.1791***	0.2354***
	(0.0489)	(0.0377)	(0.0434)	(0.0436)	(0.0444)
$\hat{\eta}_i$				0.0438	
				(0.0305)	
Observations	5033	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 columns (4) to (6), we also report corrected standard errors following Terza (2016). See the Appendix C for details; 5. The test of significance of the selection terms is a Wald test. Degrees of freedom and level of significance are in parentheses.

All in all, our opinion is that the similarities among the coefficients with and without correcting for selectivity confirm the results of our Monte Carlo experiment. A lesson for practitioners is that there is little necessity to correct for endogenous selection in situations similar to the one studied in this paper. SW's proposal is only suitable for balanced panels and after making very particular assumptions 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 have just presented in this paper.[39]

[39] To adapt the SW estimator to an unbalanced panel, we must estimate the model using the SW procedure for each subpanel (i.e., the subsamples with 4, 5, 6, 7, and so on, observations) and then recover the structural parameters by minimum distance.

Table 7: **Estimates for the dynamic log hourly earnings equation with covariates**

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Semikina Wooldridge	*No correction*	*No correction*	*Het. components correction lev eq. only*	*yby correction first dif eq*	*yby correction all equations*
	GMM	*AB*	*system*	*system*	*AB*	*system*
Lag log	0.5740***	0.1047**	0.1850***	0.1794***	0.1170***	0.2189**
hourly earnings	(0.0400)	(0.0374)	(0.0436)	(0.0442)	(0.0379)	(0.0447)
Education	0.0290***	—	0.0949***	0.0939***	—	0.0931***
	(0.004)		(0.0084)	(0.0083)		(0.0085)
Age	0.0090***	0.0070	0.0375***	0.0381**	0.0269**	0.0228***
	(0.004)	(0.0127)	(0.0113)	(0.0147)	(0.0128)	(0.0126)
Age squared	-0.0001***	-0.0001	-0.0004***	-0.0005***	-0.0001	-0.0003***
	(0.000)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0001)
$\hat{\eta}_i$				0.3020***		
				(0.0833)		
Observations	5033	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 proposal by Semikyna and Wooldridge (2013); 3. Annual dummies are included in all specifications; 4. *** significant at 1%; ** significant at 5%; * significant at 10%; 5. The standard errors have been corrected following Windmeijer (2005); In columns (4) to (6), we also report corrected standard errors following Terza (2016). See the Appendix C for details; 6. The test of significance of the selection terms is a Wald test. Degrees of freedom and level of significance are in parentheses.

6.2 Estimating models of tobacco consumption

The previous application is done on a small (cross-section dimension) sample size of $N = 550$ similar to the number of individuals of one Monte Carlo exercise. In this second application, we use a much larger (in the cross-section dimension) sample size. In more detail, we use the data in Jones and Labeaga (2003) to estimate (as they do) rational addiction models of tobacco consumption, but we also adjust a myopic model where only the lag of consumption and the price of tobacco enter the outcome equation (the price of tobacco does not enter the selection equation). 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 initial sample we use the balanced panel. The original size is 48,800 observations $N = 6100$ and $T = 8$. We drop non-smokers households, i.e., those reporting zero consumption in the eight quarters ($N = 1957$) to compare with the results of Jones and Labeaga (2003). Those households who report some zero purchases on tobacco may be affected by selection reflecting an intermittent

sequence of quits and take-ups from smoking. Then, the final model with and without correction is estimated on a sample of $N = 4041$ ($NT = 22520$), out of which 52 percent report eight positive purchases.

The results for the myopic model are presented in Table **8** (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 **8** presents first-differenced IV estimates. The rest of columns in the table reproduce the results of our first application. Columns (2) and (3) present AB and system results obtained in the selected sample, but when we do not correct the consumption equation. Column (4) adds a correction for the correlation between the unobserved heterogeneous components. In column (5) of Table **8** we present a year-by-year correction for the level equations only.

Table 8: **Estimates of myopic models of tobacco consumption**

	(1)	*(2)*	*(3)*	*(4)*	*(5)*
	2SLS-IV	*No correction*	*No correction*	*Het. components correction of lev eq. only*	*yby correction of lev eq. only*
		AB	*system*	*system*	*system*
Lag real tobacco consumption	0.2149***	0.1010***	0.1274***	0.1272***	0.0815***
	(0.0295)	(0.0263)	(0.0189)	(0.0187)	(0.0179)
Real price of tobacco	-0.8023**	-1.5900***	-0.8497***	-0.8027***	-0.3980*
	(0.4055)	(0.3614)	(0.2278)	(0.2271)	(0.2362)
$\hat{\eta}_i$				19.1929***	
				(3.1241)	
Observations	22520	22520	22520	22520	22520
Joint significance selection terms					176.39 (6)
					(0.000)

Notes: 1. $N = 4041$; 2. Quarter 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 columns (4) to (6), we also report corrected standard errors following Terza (2016). See the Appendix C for details; 5. The test of significance of the selection terms is a Wald test. Degrees of freedom and level of significance are in parentheses.

The results in column (1) correspond to IV estimates of AH. The results in column 2 are estimated by GMM on the first differenced model. As usual in myopic models, we instrument lagged consumption using previous lags of consumption. Results in column 3 are obtained using system GMM. Again, we use previous lags as instruments for consumption both in the levels and in the transformed equations (see Arellano and Bover, 1995, and Blundell and Bond, 1998). In column (4) we correct the levels equation using the proposal presented in subsubsection 5.3.2, correcting only the equation in levels of the tobacco consumption model. Despite the high significance of the

correction term, the coefficients are very similar to column 3. We find differences when correcting the levels equation with time varying selection terms in column 5. However, qualitatively, the results appear to reproduce the same characteristics already commented in the previous application.

Now, we present the results of the rational addiction model, which includes lead consumption in Table **9**. Contrary to Table **8**, we now follow the rational addiction theory and we consider that both lagged and leaded consumption are endogenous and there are not adequate consumption lags or leads to instrument them, as showed in Becker *et al.*, (1994). Column (1) reports the Jones-Labeaga estimates, which are not directly comparable to our results.[40] Columns (2) and (3) present AB and system results obtained in the selected sample, but when we do not correct the consumption equation. Column (4) adds a correction to the equation in levels for the correlation between the unobserved heterogeneous components. We present the results after including year-by-year probit corrections (in column (5) we only correct the equations in first differences and in column (6) also the equation in levels) under the assumption that the errors in both equations are contemporaneously correlated.

Table 9: **Estimates of rational addiction models of tobacco consumption**

	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
	Jones Labeaga	*No correction*	*No correction*	*Het. components correction lev eq. only*	*yby correction first dif eq*	*yby correction all equations*
	GMM	*AB*	*system*	*system*	*AB*	*system*
Lag real	0.5580***	0.9444***	0.6466***	0.6376***	0.9758***	0.6454***
tobacco consumption	(0.1185)	(0.1466)	(0.0279)	(0.0278)	(0.1498)	(0.0276)
Lead real	0.4790***	0.3611***	0.4174***	0.4052***	0.3484***	0.4098***
tobacco consumption	(0.1091)	(0.0617)	(0.0263)	(0.0083)	(0.0749)	(0.0275)
Real price	-0.0360	-0.3299	0.0086	-0.0173	-0.1584	0.0062
of tobacco	(0.0507)	(0.3035)	(0.0619)	(0.0616)	(0.3279)	(0.0643)
$\hat{\eta}_i$				-1.9844**		
				(0.8362)		
Observations	14596	14596	14596	14596	14596	14596
Joint significance	–	–	–	–	14.28 (5)	7.89 (5)
selection terms					(0.014)	(0.162)

Notes. 1. $N = 4104$; 2. System GMM results obtained by 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). See the Appendix C for details; 6. The test of significance of the selection terms is a Wald test. Degrees of freedom and level of significance are in parentheses.

The system results from Jones and Labeaga (2003) that take into account censoring are close

[40] We should be aware that they use all observations including those with observed zeros corresponding to starts-quits. Moreover, they instrument lags and leads both with prices but also with lags and leads of predicted tobacco consumption obtained in reduced form tobit or symmetrically censored least squares models.

to the system results obtained in this paper, for the lag and lead of consumption (although the instruments are different) and the AB results are also more related to the AB-kind results of Jones and Labeaga (2003) not presented here. Since there is evidence of rational addiction, the results in **8** are not adequate and price of tobacco does not affect consumption in a rational addiction framework as emphasized by the mentioned authors. We can add that correcting for selection does not affect the coefficients of the lag, lead and price, in the spirit of the results of our Monte Carlo experiments. We find, as in the previous application, some minor differences among the estimates, but they do not affect the main implications of our theoretical findings and Monte Carlo results.

7 Concluding remarks

In this paper we have analyzed the properties and the consistency of classical 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 strictly needed when both equations have common covariates. In models without common covariates, 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. Note, however, specifically for dynamic models, the system GMM estimator is moderately biased regardless of the sample size. The bias is due to caused by the level orthogonality restrictions only, thereby implying that to correct the estimator we only need to correct the level equations and not the equations in first differences. Note, however, that most of the (small) bias is due to the correlation between the individual heterogeneous components in the outcome and selection equations, which suggest a simple control approach that wipes out most of the estimator bias.

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

We evaluate the finite sample performance of the classical (FE, FD and RE) as well as IV estimators (AB and system GMM) 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 do two empirical applications differing in the number of individuals observed each period. We confirm the results of the Monte Carlo study in the estimation of female earnings equations using US data and in the adjustment of tobacco consumption equations using Spanish data.

To conclude, we find that the key determinant of the necessity of sample selection corrections

a la Heckman is the presence of a common covariate and not whether the errors of the selection and outcome equations are correlated or not. We believe that our findings could be of particular relevance for practitioners in a large variety of circumstances, specially when sample selection is very complex or of unknown form, or when selection is difficult to model due to missing data problems or lack of appropriate exclusion restrictions.

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. (2013). *Econometric Analysis of Panel Data*, John Wiley and Sons, Chichester.

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

[7] Becker, G. S., Grossman, M. and Murphy K. M. (1994). 'An empirical analysis of cigarette addiction', *American Economic Review*, 84, 396-418.

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

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

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

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

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

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

[14] Cheng, T. C., and Trivedi, P. K. (2015) Attrition Bias in Panel Data: A Sheep in Wolf's Clothing? A Case Study Based on the Mabel Survey. Health Econ., 24: 1101– 1117.

[15] 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.

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

[17] 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.

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

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

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

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

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

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

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

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

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

[27] 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.

[28] 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.

[29] Knoef J. 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.

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

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

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

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

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

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

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

[37] 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.

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

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

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

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

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

[43] 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.

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

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

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

[47] Tallis G. M. (1961). 'The moment generating function of the truncated multi-normal distribution', *Journal of the Royal Statistical Society. Series B (Methodological)*, 23, 223–229.

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

[49] 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.

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

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

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

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

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

Appendix

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 either Wooldridge (1995) for the RE case and Rochina-Barrachina (1999) for the FD case.

B.1 Recap of a dynamic model

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

$$y_{it}^* = \rho y_{it-1}^* + x_{it}\beta + \alpha_i + \varepsilon_{it} \qquad for \quad t_i \quad s.t. \quad d_{it} = 1; \tag{61}$$

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

$$d_{it}^* = z_{it}\gamma + x_{it}\delta + \eta_i + u_{it} = w_{it}\pi + \eta_i + u_{it}; \qquad d_{it} = 1\left[d_{it}^* \geq 0\right] \tag{62}$$

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 (61) and (62) 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

•***A1:*** *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:[41] $\eta_i = \bar{w}_i\theta + c_i$, where c_i is a random component independent of w_i.
•***A2:*** *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 ***A1*** and ***A2*** the reduced form selection rule of (62) 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 ***A1*** (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 ***A1***). In this case ν_{it} will be a function of $u_{i0}, ..., u_{it}, c_i$, but still independent of w_i.

B.3 Correction of biases

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

Let us consider the first-differenced model:

$$\Delta y_{it} = \rho \cdot \Delta y_{it-1} + \Delta x_{it}\beta + \Delta\varepsilon_{it} \tag{63}$$

[41] Alternatively, we can use Chamberlain's (1980) approach.

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:

$$\Delta y_{it} = \rho \cdot \Delta y_{it-1} + \Delta x_{it}\beta + E\left[\Delta\varepsilon_{it} \left| w_i, d_{it} = d_{it-1} = d_{it-2} = 1\right.\right] + \Delta e_{it} \tag{64}$$

We follow Tallis (1961) to work it out $E\left[\Delta\varepsilon_{it} \left| w_i, d_{it} = d_{it-1} = d_{it-2} = 1\right.\right]$ under a 4-variant normal distribution assumption:[42]

Assumption •*A4"': The errors* $[\Delta\varepsilon_{it}, \nu_{it}, \nu_{it-1}, \nu_{it-2}]$ *are 4-variate normally distributed and independent of* w_i.

Therefore,

$$\begin{aligned} &E\left[\Delta\varepsilon_{it} \left| w_i, d_{it} = d_{it-1} = d_{it-2} = 1\right.\right] = \sigma_{\Delta\varepsilon_t, \frac{\nu_t}{\sigma_t}} \lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) \\ &+\sigma_{\Delta\varepsilon_t, \frac{\nu_{t-1}}{\sigma_{t-1}}} \lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) + \sigma_{\Delta\varepsilon_t, \frac{\nu_{t-2}}{\sigma_{t-2}}} \lambda\left(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) \end{aligned} \tag{65}$$

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

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

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

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

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 $\lambda\left(\right)$ 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 $\Phi_3\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right)$ that appears in the denominator of the $\lambda\left(\right)$ terms. Second, we will get also estimates for the two arguments of the type $\left(H_{is} - \varrho_{t,s}H_{it}\right)\Big/\left(1-\varrho^2_{t,s}\right)^{1/2}$ in the bivariate probabilities $\Phi_2\left(\right)$. Third, we will perform all the involved bivariate probabilities $\Phi_2\left(\right)$ 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

[42] 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 $(\nu_{it}, \nu_{it-1}, \nu_{it-2})$.

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 (65) becomes:

$$\begin{aligned}
&E\left[\Delta\varepsilon_{it}\left|w_i, d_{it}=d_{it-1}=d_{it-2}=1\right.\right] = \\
&\sigma_0\left\{\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) - \lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right)\right\} \\
&-\sigma_{\varepsilon_{t-1},\frac{\nu_{t,2}}{\sigma_t}}\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) + \sigma_{\varepsilon_t,\frac{\nu_{t-1}}{\sigma_{t-1}}}\lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) \\
&+\sigma_{\varepsilon_t,\frac{\nu_{it-2}}{\sigma_{t-2}}}\lambda\left(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) - \sigma_{\varepsilon_{t-1},\frac{\nu_{it-2}}{\sigma_{t-2}}}\lambda\left(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right)
\end{aligned} \tag{66}$$

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 (66) becomes:

$$\begin{aligned}
&E\left[\Delta\varepsilon_{it}\left|w_i, d_{it}=d_{it-1}=d_{it-2}=1\right.\right] = \\
&\sigma_0\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) - \sigma_0\lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) \\
&-\sigma_{+1}\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) + \sigma_{-1}\lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) \\
&+\sigma_{-2}\lambda\left(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) - \sigma_{-1}\lambda\left(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) = \\
&\left(\sigma_0 - \sigma_{+1}\right)\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) - \left(\sigma_0 - \sigma_{-1}\right)\lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) \\
&+\left(\sigma_{-2} - \sigma_{-1}\right)\lambda\left(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right)
\end{aligned} \tag{67}$$

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 (67) becomes:

$$\begin{aligned}
&E\left[\Delta\varepsilon_{it}\left|w_i, d_{it}=d_{it-1}=d_{it-2}=1\right.\right] = \\
&\bar{\sigma}\left\{\lambda\left(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right) - \lambda\left(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right)\right\} \\
&+\bar{\sigma}_{-2}\lambda\left(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}\right)
\end{aligned} \tag{68}$$

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 in 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

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

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

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

Therefore, the corrected outcome equation (65) becomes:

$$E\left[\Delta\varepsilon_{it}\left|w_i, d_{it}=d_{it-1}=d_{it-2}=1\right.\right] = \sigma_{(\varepsilon_t),\frac{\nu_t}{\sigma_t}}\lambda\left(H_{it}\right) - \sigma_{(\varepsilon_{t-1}),\frac{\nu_{t-1}}{\sigma_{t-1}}}\lambda\left(H_{it-1}\right) \tag{69}$$

and the model simply 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 (69) becomes $\sigma_0 \{\lambda(H_{it}) - \lambda(H_{it-1})\}$.

B.3.2 Correction of the level equations

let consider now the estimation of the levels equations.

$$\begin{aligned} y_{it} &= \rho y_{it-1} + x_{it}\beta + \bar{z}_i\psi + E[\omega_{it}|z_i, d_{it} = d_{it-1} = d_{it-2} = 1] + e_{it} = \\ &\rho y_{it-1} + x_{it}\beta + \bar{z}_i\psi + \sigma_{\omega_t, \frac{\nu_t}{\sigma_t}}\lambda(H_{it}, H_{it-1}, \varrho_{t,t-1}) + \sigma_{\omega_t, \frac{\nu_{t-1}}{\sigma_{t-1}}}\lambda(H_{it-1}, H_{it}, \varrho_{t,t-1}) + e_{it} \end{aligned}, \tag{70}$$

Assumption ***•A4":*** *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 (70) becomes:

$$y_{it} = \rho y_{it-1} + x_{it}\beta + \bar{w}_i\psi + \sigma_0\lambda(H_{it}, H_{it-1}, \varrho_{t,t-1}) + \sigma_{-1}\lambda(H_{it-1}, H_{it}, \varrho_{t,t-1}) + e_{it} \tag{71}$$

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 (70) becomes:

$$\begin{aligned} y_{it} &= \rho y_{it-1} + x_{it}\beta + \bar{w}_i\psi + E[\nu_{it}|z_i, d_{it} = 1] + e_{it} = \\ &\rho y_{it-1} + x_{it}\beta + \bar{w}_i\psi + \sigma_0\lambda(H_{it}) + e_{it} \end{aligned} \tag{72}$$

and we come back again 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 allowing 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. In case of having enough sample to obtain the first difference estimator we can improve efficiency by combining the moment conditions coming from the levels and the first-differenced equations by System-GMM estimation (Arellano and Bover, 1995; Blundell and Bond, 1998).

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, (69) can be estimated with the *xtabond* Stata GMM command. In the more general stationary only case (68) can be estimated with a modified version of the *xtabond* command allowing for the two regressors: $\{\lambda(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) - \lambda(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})\}$, $\lambda(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})$. With System-GMM estimation, and under stationarity only, we could think about joint estimation with (72) and (69) with the *xtdpdsys* Stata System-GMM command if we restrict the level sample in the same way than 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 (72)) than in the equation in time differences ($\sigma_{\varepsilon_t, \frac{\nu_t}{\sigma_t}}$ in (69)).

Under *Simplification 1,* it will be more difficult to adapt standard software because, in addition to adding different regressors to the levels ($\{\lambda(H_{it}, H_{it-1}, \varrho_{t,t-1}), \lambda(H_{it-1}, H_{it}, \varrho_{t,t-1})\}$) and the differences equations ($\{\lambda(H_{it}, H_{it-1}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2}) - \lambda(H_{it-1}, H_{it}, H_{it-2}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})\}$,

$\lambda(H_{it-2}, H_{it}, H_{it-1}, \varrho_{t,t-1}, \varrho_{t,t-2}, \varrho_{t-1,t-2})$, we have to allow for different parameters associated to the sample selection correction terms in level and differenced equations.

B.5 Semiparametric model estimation

B.5.1 Correction of level equations

Consider the level model:

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

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

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

Errors can depend on the w_i only through these indices (what is called a "double index" assumption). Now (70) will become

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

The unknown function $\varphi_{jit,t-1}$, once the selection indices size has been reduced by a normal, logistic or the Heckman's lambda (inverse Mill's ratio) transformation of the selection indices, is approximated non-parametrically by a polynomial of degree q on the transformation of the indices H_{it}, H_{it-1}.[43] We could estimate the first step also by probits or a semiparametric method for binary choice with panel data.

[43] 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 (65) we could have written $E\left[\Delta\varepsilon_{it} \left| w_i, d_{it} = d_{it-1} = d_{it-2} = j\right.\right] = \varphi_{jt,t-1,t-2}\left(H_{it}, H_{it-1}, H_{it-2}\right) = \varphi_{jit,t-1,t-2}$, 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 (64) will become $\Delta y_{it} = \rho \Delta y_{it-1} + \Delta x_{it}\beta + \varphi_{jit,t-1,t-2} + \Delta e_{it}$. The unknown function $\varphi_{jit,t-1,t-2}$, once the selection indices size has been reduced by a normal, logistic or the Heckman's lambda (inverse Mill's ratio) transformation of the selection indices, is approximated non-parametrically by a polynomial of degree q on the transformation of the indices $H_{it}, H_{it-1}, H_{it-2}$.[44] We could estimate the first step also by probits or 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 The variance of corrected estimators

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

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

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

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

with

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

We choose θ which minimises:

[44] In the general case of absence of stationarity, we will interact the terms of the polynomial with time-triples dummies.

$$q = m(\theta)' \, Am(\theta)$$

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

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

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

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

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

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

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

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

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

calculate C.

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

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

Notes.

1. 1000 simulations.

2. Static selection model (A).

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

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

ÚLTIMOS DOCUMENTOS DE TRABAJO

2020-06: "Consistent estimation of panel data sample selection models", **Sergi Jiménez-Martín, José M. Labeaga y Majid al Sadoon**.
2020-05: "Encajando el puzle: Una estimación rápida del número de infectados por COVID-19 en España a partir de fuentes indirecta", **David Martín-Barroso, Juan A. Núñez-Serrano, Jaime Turrión y Francisco J. Velázquez**.
2020-04: "Does the Liquidity Trap Exist?, **Stéphane Lhuissier, Benoît Mojon y Juan Rubio-Ramírez**.
2020-03: "How effective has been the Spanish lockdown to battle COVID-19? A spatial analysis of the coronavirus propagation across provinces", **Luis Orea e Inmaculada C. Álvarez**.
2020-02: "Capital humano y crecimiento: teoría, datos y evidencia empírica", **Angel de la Fuente**.
2020-01: "Twin Default Crises", **Caterina Mendicino, Kalin Nikolov, Juan Rubio-Ramirez y Javier Suarez**.
2019-07: "Vivienda y política pública: objetivos e instrumentos", **Miguel-Ángel López García**.
2019-06: "Mercados, entidades financieras y bancos centrales ante el cambio climático: retos y oportunidades", **Clara I. González y Soledad Núñez.**
2019-05: "Education and Gender Differences in Mortality Rates", **Cristina Belles-Obrero, Sergi Jiménez-Martín y Judit Vall Castello.**
2019-04: "Las viviendas turísticas ofertadas por plataformas on-line: Estado de la cuestión", **Armando Ortuño y Juan Luis Jiménez**.
2019-03: "Now-casting Spain", **Manu García y Juan F. Rubio-Ramírez**.
2019-02: "Mothers' care: reversing early childhood health shocks through parental investments", **Cristina Belles-Obrero, Antonio Cabrales, Sergi Jimenez-Martin y Judit Vall-Castello**.
2019-01: "Measuring the economic effects of transport improvements", **Ginés de Rus y Per-Olov Johansson**.
2018-15: "Diversidad de Género en los Consejos: el caso de España tras la Ley de Igualdad",**J. Ignacio Conde-Ruiz, Manu García y Manuel Yáñez**.
2018-14: "How can urban congestion be mitigated? Low emission zones vs. congestion tolls", **Valeria Bernardo, Xavier Fageda y Ricardo Flores-Fillol**.
2018-13: "Inference in Bayesian Proxy-SVARs", **Jonas E. Arias, Juan F. Rubio-Ramírez y Daniel F. Waggoner**.
2018-12: "Evaluating Large Projects when there are Substitutes: Looking for Possible Shortcuts", **Per-Olov Johansson y Ginés de Rus**.
2018-11: "Planning, evaluation and financing of transport infrastructures: Rethinking the basics", **Ginés de Rus y M. Pilar Socorro**.
2018-10: "Autonomía tributaria subnacional en América Latina", **Juan Pablo Jiménez e Ignacio Ruelas**.
2018-09: "Ambition beyond feasibility?Equalization transfers to regional and local governments in Italy", **Giorgio Brosio.**
2018-08: "Equalisation among the states in Germany: The Junction between Solidarity and Subsidiarity", **Jan Werner.**
2018-07: "Child Marriage and Infant Mortality: Evidence from Ethiopia", **J.García-Hombrados**
2018-06: "Women across Subfields in Economics: Relative Performance and Beliefs", **P. Beneitoa, J.E. Boscá, J. Ferria y M. García**.
2018-05: "Financial and Fiscal Shocks in the Great Recession and Recovery of the Spanish Economy", **J. E. Boscá, R. Doménech, J. Ferri, R. Méndez y J. F. Rubio-Ramírez**.
2018-04: "Transformación digital y consecuencias para el empleo en España. Una revisión de la investigación reciente", **Lucas Gortazar**.
2018-03: "Estimation of competing risks duration models with unobserved heterogeneity using hsmlogit", **David Troncoso Ponce**.
2018-02: "Redistributive effects of regional transfers: a conceptual framework", **Julio López-Laborda y Antoni Zabalza**.
2018-01: "European Pension System: ¿Bismarck or Beveridge?", **J. Ignacio Conde-Ruiz y Clara I. González**.
2017-15: "Estimating Engel curves: A new way to improve the SILC-HBS matching process", **Julio López-Laborda, Carmen Marín-González y Jorge Onrubia.**
2017-14: "New Approaches to the Study of Long Term Non-Employment Duration in Italy, Germany and Spain", **B. Contini, J. Ignacio Garcia Perez, T. Pusch y R. Quaranta**.
2017-13: "Structural Scenario Analysis and Stress Testing with Vector Autoregressions", **Juan Antolín-Díıaz y Juan F. Rubio-Ramírez**.
2017-12: "The effect of changing the number of elective hospital admissions on the levels of emergency provision", **Sergi Jimenez-Martin, Catia Nicodemo y Stuart Redding**.
2017-11: "Relevance of clinical judgement and risk stratification in the success of integrated care for multimorbid patients", **Myriam Soto-Gordoa, Esteban de Manuel, Ane Fullaondo, Marisa Merino, Arantzazu Arrospide, Juan Ignacio Igartua y Javier Mar**.
2017-10: "Moral Hazard versus Liquidity and the Optimal Timing of Unemployment Benefits", **Rodolfo G. Campos, J. Ignacio García-Pérez y Iliana Reggio**.
2017-09: "Un análisis de modelos para financiar la educación terciaria: descripción y evaluación de impacto", **Brindusa Anghel, Antonio Cabrales, Maia Güell y Analía Viola**.
2017-08: "Great Recession and Disability Insurance in Spain", **Sergi Jiménez-Martín, Arnau Juanmarti Mestres y Judit Vall Castelló**.
2017-07: "Narrative Sign Restrictions for SVARs", **Juan Antolín-Díaz y Juan F. Rubio-Ramírez**.
2017-06: "Faster estimation of discrete time duration models with unobserved heterogeneity using hshaz2", **David Troncoso Ponce**.