Documento de Trabajo - 2016/01

Monte Carlo evidence on the estimation of AR(1) panel data sample selection models

Sergi Jiménez-Martín
Universitat Pompeu Fabra, Barcelona GSE and FEDEA

José María Labeaga
UNED and UNU-MERIT, University of Maastricht

January 2016

fedea

Monte Carlo evidence on the estimation of AR(1) panel data sample selection models*

Sergi Jiménez-Martín, Universitat Pompeu Fabra, Barcelona GSE and FEDEA[a]
José María Labeaga, UNED and UNU-MERIT, University of Maastricht

November, 2015

Abstract

We present Generalized Method of Moments estimators for AR(1) dynamic panel data sample selection models. We perform a Monte Carlo study to evaluate the finite sample properties of the proposed estimators. Our results suggest that correcting for sample selection in many standard cases does not add much to the uncorrected estimates. In particular, the magnitude of the biases is similar and very small when estimating the model either correcting or not the equation of interest. This equivalence also holds in the dynamic model with exogenous regressors. These results are especially relevant for practitioners either when there is selection of unknown form or selection is difficult to model.

Keywords: Panel data, sample selection, dynamic model, generalized method of moments
JEL Class.: J52, C23, C24

* We are grateful to ECO2011-30323-C03-02, ECO2014-52238-R, and ECO2012-39553-C04-01 (Spanish Ministry of Economy and Competitiveness) for financial support. We are also grateful to Badi Baltagi and María Rochina-Barrachina for some very useful comments and to participants at the 2015 Annual Meeting of the International Applied Econometric Society held at Thessaloniki, especially to Frank Windmeijer and Juan M. Rodríguez-Poo. Any remaining error is our responsibility.
[a] *Corresponding author:* Sergi Jiménez-Martín, Department of Economics, UPF, Barcelona, Spain. e-mail: sergi.jimenez@upf.edu

1. Introduction

The increasing availability of longitudinal data has provided the possibility of doing both theoretical and empirical papers in several economic fields. As it is well known, panel data offers researchers some advantages both with respect to cross-section and time series. The main advantage is that panel data methods can account for unobserved heterogeneity. However, while in linear models it is normally easy to estimate the parameters even under the presence of correlated unobserved heterogeneity, the same is not true in the case of non-linear models.The problems of self-selection, non-response and attrition are usually worse in panels than in cross-sections (see Baltagi, 2005). In many empirical applications, these problems entail the necessity to estimate the models on unbalanced panels. Many times, we should first answer the question about the reason why the panel becomes unbalanced and it is quite common for it to appear because of endogenous attrition or endogenous selection.

There are a number of studies dealing at the same time with unobserved heterogeneity and selectivity. Most of them do it under strict exogeneity assumptions, such as Verbeek and Nijman (1992) who proposed tests of selection bias either with or without allowance for correlation between the unobserved effects and explanatory variables. Wooldridge (1995) also proposed variable addition tests for selection bias and he gives procedures for estimating the model after correcting for selectivity. Kyriazidou (1997) proposes correcting for selection bias by using a semiparametric approach based on a conditional exchangeability assumption. Vella and Verbeek (1998) allowed for endogenous explanatory variables in the outcome equation. Rochina-Barrachina (1999) proposed estimators where the correction terms are more complex than in Wooldridge (1995) because the model is estimated in time differences. Kyriazidou (2001) extends her previous methods to dynamic models with selection while Hu (2002) constitutes an example for the case of dynamic censored panel data models with a lagged latent dependent variable. Finally, Semykina and Wooldridge (2010, 2013) propose new two-stage methods for estimating panel data models in the presence of endogeneity, dynamics and selection.[1]

[1] More recent theoretical papers have explored either bias bias-corrected estimators for the static case (Fernández-Val and Vella, 2007), semiparametric (Gayle and Viauroux, 2007, Sasaki, 2015), or maximum likelihood estimators (Raymond *et al*., 2010 or Lai and Tsay, 2012) for the dynamic case. In contrast to these proposals, our aim is to provide solutions easy to apply from the point of view of an applied practitioner.

The different methods in the previous papers have been applied to a number of empirical studies. Charlier, Melenberg and van Soest (2001) apply them to estimate housing expenditure by households. Jones and Labeaga (2004) select out the sample of non-smokers using the variable addition tests of Wooldridge (1995) and then estimate tobit-type models on the sample of smokers and potential smokers using Generalized Method of Moments (GMM) and Minimum Distance (MD) methods. González-Chapela (2007) uses GMM when estimating the effects of recreation goods on male labour supply. Winder (2004) uses instrumental variables to account for endogeneity of some regressors when estimating earnings equations for females. Jiménez-Martín (2006) estimates and tests the possibility of different wage equations for strikers and non-strikers in a dynamic context. Dustmann and Rochina-Barrachina (2007) estimate females' wage equations extending Rochina-Barrachina (1999). Finally, Semykina and Wooldridge (2010, 2013) apply their methods to estimate earnings equations for females.

Since it is likely for these approaches to be used more frequently in the future, we think that it is important to highlight advantages and problems in the performance of the different estimators and to draw researchers' attention to potential pitfalls in using them in applied studies. In particular, in this paper we focus on the estimation of the AR(1) dynamic panel data sample selection model. We assume a typical model for the outcome of interest and we allow the selection equation to be either static or dynamic. We also assume a two error component in both equations with a very general correlation structure. This model is then evaluated using Monte Carlo methods under different assumptions. The correction for selectivity is based on estimates resulting in typical binomial probit models adjusted for each cross-section. The corrected outcome equation is then estimated using a system GMM estimator that can be implemented with standard software.

This exercise provides a general picture implying little need to correct for selectivity when we allow for moderate (or even high) degrees of selection. Our results also apply to outcome equations with exogenous regressors. Analysis to test their sensitivity to different maintained assumptions also show that they are very robust except for the case where the ratio of variances of the heterogeneus component to the idyosincratic error is high. We find that these results could be especially relevant for practicioners in those cases in which there is selection of unknown form or selection is difficult to model.

In section 2 we present the general model and the estimation methods. The performance of the proposed estimators is tested in section 3 where we present a Monte Carlo study of the finite sample average bias of GMM estimators as well as a sensitivity analysis to some maintained assumptions. Section 4 concludes.

2. The model

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

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

for *i = 1, ..., N* and *t = 1, ..., T.* α_i is an individual heterogeneous component independent of the idiosyncratic error ε_{it}, and ρ a parameter to estimate. 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:

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

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

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

in a way such that $d_{it} = 1$ if y_{it} is observed and zero otherwise. The selection equation (2) could also contain a lagged observed indicator (d_{it-1}) which we ignore for the moment to keep notation as simple as posible.

In the absence of selection and for the typical situation of *N* large and *T* small, model (1) in first differences is usually estimated by instrumental variables (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 GMM approach to include equations in levels and proposed the estimation of the whole model using system GMM. As noted by Blundell and Bond (1998) in the case of an AR(1) with highly persistent series first-differencing could lead to a weak instruments problem. Then, the use of equations in levels becomes important to improve efficiency.

2.1. Estimation of the outcome equation under selection

Due to the fact that simple methods (least squares, within-groups) do no work for the dynamic model, an easy alternative for practicioners that also control unobserved heterogeneity (and any potential correlation of the time invariant component) is to estimate (1) subject to (2) in the first-differenced model. First differences introduces serial correlation in ε_{it}, so we have to use IV. In this pure autoregressive model, the best alternative is the use of internal instruments and in first differenced models we have to use instruments lagged at least twice. The sample is conditional to observing the outcome for at least three consecutive periods $d_{it} = d_{it-1} = d_{it-2} = 1$ and the amount of data lost depends on the degree of selection. As suggested by Blundell and Bond (1998), to improve both efficient and small sample consistency of the IV estimator we opt for using the system GMM method.[2]

For the system GMM method the estimating sample differs by equation when the instruments consist in lagged dependent variables. For the levels equations we have:

$$y_{it} = \rho y_{it-1} + \alpha_i + E(\varepsilon_{it} \,/\, z_{it}, d_{it} = d_{it-1} = 1)$$

for observations such that $d_{it} = d_{it-1} = 1$. And for the first differenced equations we have:

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

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

Since GMM methods are based on instruments that are uncorrelated with both the errors in levels ε_{it} and in first differences $\Delta\varepsilon_{it}$, it should be feasible to recover consistent estimates of the parameters of the model. For the first differences equations all the values of y lagged at least twice are valid instruments. In addition to them, for the levels equations, Δy_{it-1} is also valid. Note, that in order to construct a valid instrument for the levels equation we need to condition the sample on three consecutive positive outcomes ($d_{it} = d_{it-1} = d_{it-2} = 1$),

[2] Since our simulation results show that system GMM estimates outperform GMM first-differences estimates, we develop the analysis based on system GMM. However, GMM first-differences estimates are available on request.

making the effective sample condition identical for both level and first differences equations.

Moreover, as the final estimating sample is selected on positives for at least three consecutive previous periods, we feel we will not have much necessity of correcting the bias.[3] Yet, we consider two alternatives to correct it. First, we will use a very simple method for the static selection model in (2). Second, we will model the heterogeneity using the proposal of Chamberlain (1984).

2.2. Different approaches to correction

For a typical selection model (2) and assuming normality of $\eta_i + u_{it}$ we can estimate a probit for each period and then compute the well known selection term $\hat{\lambda}_{it}(z_{it}\hat{\gamma})$. In a second step, equation (1) can be estimated (combining equations in levels and first differences) introducing $\hat{\lambda}_{it}(z_{it}\hat{\gamma})$ in levels for the equations in levels and first differenced for the equations in differences. When we allow correlation between z_{it} and η_i we can rely on Mundlak (1978) and assume $\eta_i = \bar{z}_{i.}\varphi$. Again, we can estimate a probit for each period and compute $\tilde{\lambda}_{it}(z_{it}\tilde{\gamma} + \bar{z}_{i.}\tilde{\varphi})$, which is introduced in a second step as before. In the case of a dynamic selection equation, the lagged regressor is correlated with the random effect by construction and we need to rely either on Mundlak's proposal or on a less restrictive one due to Chamberlain (1984). In the latter case, we can assume $\eta_i = \pi_1 z_{i1} + \pi_2 z_{i2} + \cdots + \pi_T z_{iT}$ and recover the corresponding selection terms.[4] When the selection equation is dynamic we follow this last alternative and for coherence we also implement it for the static selection model. However, given our (simplifying) assumptions about z_{it} the three alternatives we have described here lead to statistically identical results for the static selection case.

---

[3] Arellano *et al.* (1999) proposed the estimation of sample selection models conditioning on exogenous positive past outcomes and they show that the degree of selection is reduced by significant proportions in economic models with persistence. In our case, the introduction of the correction terms implies the selection of the estimating sample and it seems that the consequences could be similar.

[4] Strictly speaking, in order to recover the structural parameters of the selection equation we should estimate a probit for each year based on a reduced form where d_{it}^* is modeled as a function of all exogenous variables (the *z*'s) and we predict the index $\hat{d}_{it}^*$. Then, in a second stage we estimate the structural parameters by MD or GMM.

We can still use a more general correction (see Jiménez *et al.*, 2009). Under a fairly standard stationarity condition of the selection process, these authors find that estimation gets more complicated since the correction of equations requires additional regressors based on bivariate probit estimates for periods t and t-1. Specifically, if we name $\widehat{H}_{it} = f(z)$, with $f(z)$ a function of the exogenous variables, the estimation in levels requires the terms $\hat{\lambda}_{it}(\widehat{H}_{it}, \widehat{H}_{it-1}, \hat{\rho}_{t,t-1})$ and $\hat{\lambda}_{it}(\widehat{H}_{it-1}, \widehat{H}_{it}, \hat{\rho}_{t,t-1})$. Likewise, equations in first differences must include selection terms obtained from a trivariate probit estimated for periods t, t-1 and t-2 (for details see Rochina-Barrachina, 1999 or Jiménez *et al.*, 2009).

3. Simulation study

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

$$d^*_{it} = a - z_{it} - \eta_i - u_{it} \qquad (4.1)$$

$$d^*_{it} = a - 0.5d_{it-1} + z_{it} - \eta_i - u_{it} \qquad (4.2)$$

$$d_{it} = 1[d^*_{it} > 0] \qquad (4.3)$$

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:

$$y^*_{it} = (2 + \alpha_i + \varepsilon_{it})/(1 - \rho) \text{ if } t = 1 \qquad (5.1)$$

$$y^*_{it} = 2 + \rho y^*_{it-1} + \alpha_i + \varepsilon_{it} \text{ if } t = 2, \ldots, T \qquad (5.2)$$

$$y_{it} = y^*_{it} \text{ if } d_{it} = 1 \qquad (5.3)$$

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

$$\eta_i \sim N(0, \sigma_\eta) \text{ with } \sigma_\eta = 1 \qquad (6.1)$$

$$u_{it} \sim N(0, \sigma_u) \text{ with } \sigma_u = 1 \qquad (6.2)$$

$$\alpha_i = \alpha^0_i + 0.5\eta_i, \alpha^0_i \sim N(0, \sigma_{\alpha^0}) \text{ with } \sigma_{\alpha^0} = 1 \qquad (6.3)$$

$$\varepsilon_{it} = \varepsilon_{it}^{0} + 0.5u_{it},\ \varepsilon_{it}^{0} \sim N(0, \sigma_{\varepsilon^0}) \text{ with } \sigma_{\varepsilon^0} = 1 \qquad (6.4)$$

These assumptions imply that $corr(\varepsilon_{it}, u_{it}) = corr(\alpha_i, \eta_i) = 0.5\sqrt{1 + 0.5^2} = 0.4472$.

3.1. Description of the experiments

For each experiment, *N* = 500 and for each *i* we draw up to 20 times series observations. In order to take care of initial conditions we end up having a small T (from 4 to 7) as it is usual in the empirical literature. Once selection is applied the unbalanced panels are formed. At least three consecutive observations of the same regime are needed in order to form an observation of the selected panel. For each combination of the parameters we perform 500 replications.

The structure of the model makes selection of the instruments a crucial step of this simulation study. We select them as follows: for first difference equations we use lags from *t-2* backwards, although we also evaluate the performance of the estimates with a restricted set of instruments. For the equation in levels we use the lagged first-difference of the outcome as an additional instrument. 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 remaining for estimation.

We present three alternatives of the system GMM estimator to compare the performance of the model under different assumptions. The first set of estimates is obtained under the assumption of exogenous selection (i.e Utility interdependence and consumption behavior: the roles of envy and habits, no correlation between the time varying errors is impossed). The other two are obtained under the assumption of endogenous selection but one of them does not correct for it. We can see these results as estimates on the positives (adequately selected) and they evaluate the necessity of adding a selection term. The last one corrects using the selection corrections obtained after estimating the index equation with a year-by-year probit (Wooldridge, 1995), accounting for correlated effects when necessary.

3.2. Simulation results for the pure autoregressive model

Table 1 presents results for AR(1) model for three values of the autoregressive parameter:

0.25, 0.50 and 0.75.[5] We simulate two alternative selection models as presented in equations (4.1) and (4.2). For each combination of selection model and autoregressive parameter we report three system GMM estimators constructed under different assumptions about the sampling process: (a) no endogenous selection; (b) endogenous selection without correction and (c) endogenous selection with a year by-year probit correction for selection bias. We also present the average significance and the empirical rejection frequency of the variable addition test for $\lambda(.)$.

The results without correction have very strong implications. In all cases and for both models, the bias never exceeds one percent (average bias divided by the true value of the coefficient) and in many cases it is even smaller (close to zero). Adding a simple correction (based on a year-by-year probit) reduces the bias by about 10 percent and also reduces the standard errors of the model. However, this is not very crucial when the bias is relatively small.

The average significance test of the null hypothesis that the coefficient of $\hat{\lambda}_{it}(.) = 0$ is estimated at about 0.02 for the static selection equation and 0.03 for the dynamic one. Alternatively, the implied rejection frequency is estimated at about 0.90 and 0.86 respectively for both models, somewhat below the expected value of 0.95. In summary, the evidence seems to suggest little need to correct for sample selection bias in pure autoregressive dynamic panel models when the degree of selection is moderate. Since selection is based on variables not correlated with the outcome, the inclusion of $\hat{\lambda}_{it}(.)$ in a model such as (1) does not affect the bias in estimating ρ. Just as an example, in a simulation taken at random the covariance between $\hat{\lambda}_{it}(.)$ and y_{it-1} corresponding to selection model A of Table 1 is -0.036 (-0.058) for ρ=0.25 (ρ=0.75) and the variance of y_{it-1} is 5.56 (21.2) so the bias is less than 0.65 percent (0.27 percent) for plausible values of the parameter of the selection variable.

[Insert Table 1 around here]

3.3. Sensitivity analysis

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

In this section we comment on various departures from the basic set of assumptions. In particular, we consider the following representative cases: (a) varying the longitudinal dimension; (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 sample size to $N = 200$; (e) reducing the correlation between the errors. We only present the results in Table 2 for the case of an autoregressive coefficient $\rho = 0.25$.[6]

Our first experiment to test sensitivity varies the longitudinal dimension of the panel from $T = 7$ to $T = 4$. Apart from the expected increase in the estimated variance, the effect on average bias of the AR(1) coefficient implied by this change is almost negligible. When the time series dimension of the panel varies from 4 to 7, the results lie in between those presented in Tables 1 and 2. Alternatively, the significance of the test for $\hat{\lambda}_{it}(.) = 0$ gets reduced, so we are more likely to accept there is no relevant selection.

Increasing the degree of sample selection from 0.15 to 0.25 reduces, in general, the average bias of the autoregressive coefficient, especially for values of ρ larger than 0.25 (that are not shown in Table 2, but that are available on request). In addition, it mildly increases its variance due to the significant reduction in the number of observations selected in the sample (the average number of observations reduces about 30 percent). Finally, both the significance and the estimated rejection frequency of the variable addition test increase significantly (especially for the static selection model) and now the test clearly detects endogenous selection. Our guess is that a larger fraction of zeroes in d_{it}^* helps to identify $\hat{\lambda}_{it}(.)$ but since its correlation with y_{it-1} is small, so is the bias.

When the ratio of variances of the outcome equation increases as the individual heterogeneity variance is double the time series variance, identification of the autoregressive parameter becomes harder without endogenous selection. Alternatively, it does not change so much with endogenous selection. Furthermore, the estimated rejection frequency of the test of the coefficient of the correction term does not improve significantly. Further increases of the ratio of the variances increase the bias of the system GMM estimator even under exogenous selection of the sample.

[6] Results for other values of the autoregressive parameter are available on request.

[Insert Table 2 around here]

As in Blundell and Bond (1998), having a small cross-section does not have important implications for the results of the experiment. Naturally, we get larger standard error than before due to the important reduction of the sample size.

Finally, we consider a reduction in the correlation of the time varying errors. In particular we assume the following structure for the errors: $\varepsilon_{it} = \varepsilon_{it}^{0} + 0.25u_{it}$, which implies a correlation coefficient of 0.2425 (=$0.25/\sqrt{1+0.25^2}$). As it can be easily detected comparing the results reported in Tables 2 and 1, this change has no significant impact on the average bias of the various estimators we have considered.

4. Concluding remarks

In this paper we have analyzed the performance of GMM estimators of an AR(1) panel data model subject to sample selection from the point of view of practicioners. In particular we evaluate the performance of the estimator in three situations: no endogeneous selection, and endogenous selection controlling and not controlling for sample selection. To see the performance of the proposed estimator we perform a Monte Carlo study of the finite sample properties of the proposed methods.

Our Monte Carlo results suggest that in many standard cases there is little need to correct for sample selection. This is true in general for the purely autoregressive model and the dynamic model with exogenous regressors. Analysis to test the sensitivity of the results to different maintained assumptions also show that they are very robust except for the case where the ratio of the variances of the heterogeneus component to the idyosincratic error is very high. However, in the latter case the bias is not worse than the one obtained in absence of sample selection.

In summary, we believe that these results could be especially relevant for practicioners in those cases in which there is selection of unknown form or selection is difficult to model.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

Lai, H. P. and Tsay, W. J. (2012). 'Maximum likelihood estimation of the panel data sample selection model', IEAS Working Paper 12-A006.

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

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

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

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

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

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

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

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

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

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

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

Table 1. Average bias in the AR(1) model

			No endogenous selection	Endogenous selection			
				No correction	Year by year correction	Bias correction testing	
rho	selection[1] model	stat	bias	bias	bias	average[2] signif	ERF[3]
(0.25)	A	av.	-0.0020	0.0003	0.0000	0.016	0.902
		s.e.	*0.0427*	*0.0426*	*0.0420*		
(0.50)	A	av.	-0.0027	0.0052	0.0046	0.016	0.908
		s.e.	*0.0503*	*0.0505*	*0.0499*		
(0.75)	A	av.	-0.0075	0.0072	0.0063	0.018	0.906
		s.e.	*0.0618*	*0.0637*	*0.0631*		
(0.25)	B	av.	-0.0017	-0.0011	-0.0021	0.027	0.856
		s.e.	*0.0431*	*0.0429*	*0.0424*		
(0.50)	B	av.	-0.0024	0.0035	0.0022	0.027	0.856
		s.e.	*0.0506*	*0.0508*	*0.0503*		
(0.75)	B	av.	-0.0072	0.0057	0.0041	0.028	0.864
		s.e.	*0.0616*	*0.0627*	*0.0622*		

Notes.

1. A: static selection as in (4.1). B: dynamic selection as in (4.2).
2. Average signif. = $E\left[1 - \Phi\left(|\frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}}|\right)\right]$ with $\Phi(.)$ being the standard normal cdf, $\hat{\theta}$ the coefficient of the correction term $\hat{\lambda}_{it}(.)$ and $\hat{\sigma}_{\hat{\theta}}$ its standard error.
3. ERF = empirical rejection frequency = $1 - \Phi\left(|\frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}}|\right)$ < = 0.05.
4. Sample size N = 500. Number of replications = 500.
5. All results are obtained using the system GMM estimator.

Table 2. Sensitivity analysis: average bias under alternative scenarios

			No endogenous selection	Endogenous selection			
				No correction	Year by year correction	Bias correction testing	
rho	selection model	stat	bias	bias	bias	average signif	ERF
1. Decreasing the max longitudinal dimension to T = 4							
(0.25)	A	av.	-0.0056	0.0031	0.0035	0.070	0.654
		s.e.	*0.0732*	*0.0762*	*0.0769*		
(0.25)	B	av.	-0068	-.0060	-.0073	0.090	0.578
		s.e.	*0.0724*	*0.0764*	*0.0763*		
2. Average sample selection increased to 0.25							
(0.25)	A	*av.*	-0.0001	0.0008	-0.0013	0.007	0.962
		s.e.	*0.0505*	*0.0503*	*0.0498*		
(0.25)	B	*av.*	-0.0010	-0.0033	-0.0053	0.014	0.920
		s.e.	*0.0505*	*0.0509*	*0.0506*		
3. Increasing the ratio of the variances ($\sigma_\alpha / \sigma_\varepsilon = 2$)							
(0.25)	A	*av.*	-0.0043	0.0020	0.0022	0.021	0.870
		s.e.	*0.0463*	*0.0451*	*0.0445*		
(0.25)	B	*av.*	-0.0038	-0.0032	-0.0041	0.031	0.826
		s.e.	*0.0468*	*0.0453*	*0.0448*		
4. Reduced sample size (N = 200)							
(0.25)	A	*av.*	0.0030	0.0021	0.00153	0.109	0.518
		s.e.	*0.0732*	*0.0677*	*0.0680*		
(0.25)	B	*av.*	0.0028	0.0002	-0.0009	0.126	0.462
		s.e.	*0.0731*	*0.0682*	*0.0681*		
5. Decreasing the correlation between the errors to 0.2425							
(0.25)	A	*av.*	-0.0021	-0.0024	-0.0025	0.109	0.486
		s.e.	*0.0425*	*0.0436*	*0.0433*		
(0.25)	B	*av.*	-0.0016	-0.0024	-0.0028	0.126	0.446
		s.e.	*0.0429*	*0.0434*	*0.0432*		

Notes.

1. A: static selection as in (4.1). B: dynamic selection as in (4.2).
2. Average signif. = $E\left[1-\Phi\left(|\frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}}|\right)\right]$ with $\Phi(.)$ being the standard normal cdf, $\hat{\theta}$ the coefficient of the correction term $\hat{\lambda}_{it}(.)$ and $\hat{\sigma}_{\hat{\theta}}$ its standard error.
3. ERF = empirical rejection frequency = $1-\Phi\left(|\frac{\hat{\theta}}{\hat{\sigma}_{\hat{\theta}}}|\right)$ <= 0.05.
4. Sample size N = 500 (except for case 4). Number of replications = 500.
5. All results are obtained using the system GMM estimator.

ÚLTIMOS DOCUMENTOS DE TRABAJO

2016-01: "Monte Carlo evidence on the estimation of AR(1) panel data sample selection models", **Sergi Jiménez-Martín y José María Labeaga**.
2015-13: "On the Treatment of Foreigners and Foreign-Owned Firms in Cost–Benefit Analysis", **Per-Olov Johansson y Ginés de Rus**.
2015-12: "Evaluating Options for Shifting Tax Burden to Top Income Earners", **Jorge Onrubia, Fidel Picos y María del Carmen Rodado**.
2015-11: "Differences in Job De-Routinization in OECD countries: Evidence from PIAAC", **Sara De La Rica y Lucas Gortazar**.
2015-10: "Bad times, slimmer children?", **Cristina Belles-Obrero, Sergi Jimenez-Martín y Judit Vall-Castello**.
2015-09: "The Unintended Effects of Increasing the Legal Working Age on Family Behaviour", **Cristina Belles-Obrero, Sergi Jimenez-Martín y Judit Vall-Castello**.
2015-08: "Capital Humano y Productividad", **Ángel de la Fuente**.
2015-07: "The effect of changes in the statutory minimum working age on educational, labor and health outcomes", **Sergi Jiménez-Martín, Judit Vall y Elena del Rey**.
2015-06: "The Effects of Employment Uncertainty, Unemployment Insurance, and Wealth Shocks on the Retirement Behavior of Older Americans", **Hugo Benítez-Silva, J. Ignacio García-Pérez y Sergi Jiménez-Martín**.
2015-05: "Instruments, rules and household debt: The effects of fiscal policy", **J. Andrés, J.E. Boscá y J. Ferri**.
2015-04: "Can International Macroeconomic Models Explain Low-Frequency Movements of Real Exchange Rates?", **Pau Rabanal y Juan F. Rubio-Ramírez**.
2015-03: "Privatización, competencia y regulación aeroportuaria: Experiencia internacional", **Ofelia Betancor y María Paz Espinosa**.
2015-02: "La experiencia internacional en alta velocidad ferroviaria", **Daniel Albalate y Germà Bel**.
2015-01: "Household Debt and Fiscal Multipliers", **J. Andrés, J.E. Boscá y J. Ferri**.
2014-21: "Structural Estimation of a Model of School Choices: the Boston Mechanism vs. Its Alternatives", **Caterina Calsamiglia, Chao Fu y Maia Güell**.
2014-20: "Which club should I attend, Dad?: Targeted socialization and production", **Facundo Albornoz, Antonio Cabrales y Esther Hauk**.
2014-19: "The Informational Content of Surnames, the Evolution of Intergenerational Mobility and Assortative **Mating", Maia Güell, José V. Rodríguez Mora y Chris Telmer**.
2014-18: "Risk-sharing and contagion in networks", **Antonio Cabrales, Piero Gottardi y Fernando Vega-Redondo**.
2014-17: "A simple model of aggregate pension expenditure", **Ángel de la Fuente**.
2014-16: "The economic evaluation of infrastructure investment. Some inescapable tradeoffs", **Ginés de Rus**.
2014-15: "Cross-country data on the quantity of schooling: a selective survey and some quality measures", **Ángel de la Fuente y Rafael Doménech**.
2014-14: "Educational Attainmet in the OECD, 1960-2010, (version 3.1)", **Ángel de la Fuente y Rafael Doménech**.
2014-13: "The Systematic Component of Monetary Policy in SVARs: An Agnostic Identification Procedure", **Jonas E. Arias, Dario Caldara y Juan F. Rubio-Ramírez**.
2014-12: "Reforming the U.S. Social Security system accounting for employment uncertainty", **Hugo Benítez-Silva, J. Ignacio García-Pérez y Sergi Jiménez-Martín**.
2014-11: "Estimating Dynamic Equilibrium Models with Stochastic Volatility", **Jesús Fernández-Villaverde, Pablo Guerrón-Quintana y Juan F. Rubio-Ramírez**.
2014-10: "Efficiency and Endogenous Fertility", **Mikel Pérez-Nievas, J. Ignacio Conde-Ruiz y Eduardo L. Giménez**.
2014-09: "The Role of Global Value Chains during the Crisis: Evidence from Spanish and European Firms", **Aranzazu Crespo y Marcel Jansen**.
2014-08: "Can Fixed-Term Contracts Put Low Skilled Youth on a Better Career Path? Evidence from Spain", **J. Ignacio García Pérez, Ioana Marinescu y Judit Vall Castello**.
2014-07: "Gender Peer Effects in School, a Birth Cohort Approach", **Antonio Ciccone y Walter Garcia-Fontes**.
2014-06: "Delaying the Normal and Early Retirement Ages in Spain: Behavioural and Welfare Consequences for Employed and Unemployed Workers", **Alfonso R. Sánchez, J. Ignacio García-Pérez y Sergi Jiménez-Martín**.
2014-05: "Immigrant Selection over the Business Cycle: The Spanish Boom and the Great Recession", **Jesús Fernández-Huertas Moraga**.
2014-04: "The Incentive Effects of Minimum Pensions: extended version", **Sergi Jiménez-Martín**.
2014-03: "A Practitioners' Guide to Gravity Models of International Migration", **Michel Beine, Simone Bertoli y Jesús Fernández-Huertas Moraga**.
2014-02: "L'auberge Espagnole y el Apartamento Francés: los Determinantes del Aprendizaje del Francés en España", **Brindusa Anghel y Maia Güell**.