A note on the estimation of AR(1) fixed-effects regressions in unbalanced panels.

Alexandre Cazenave-Lacroutz[*,a,b], Vieu Lin[*]

October 18, 2019

Version 1.0.5 - comments are welcome !

Abstract

This note discusses the estimation of the parameters of fixed-effects regression with AR(1) perturbations. In the most general case of non-random fixed-effects model, it notably highlights that current estimation procedures (based on the Baltagi et Wu (1999) transformation) fails to take into account possible missingness of the data, with two main contributions. First, it highlights that the variance of the perturbations is currently badly estimated in this most general case as soon as some observations are missing, and proposes a method to consistently estimate it. Second, it shows the Baltagi et Wu (1999) transformation is not adapted in this case, and suggests a novel transformation that produces a consistent estimation of β (that may be more efficient than estimating β by ignoring the autocorrelation of the perturbations). Monte Carlo simulations illustrate the properties of these new estimation methods.

[*] Institut National de la Statistique et des Études Économiques, 88 Avenue Verdier, Montrouge. This document does not reflect the position of Insee, Université Paris Dauphine or Crest, but only its authors' views. This research was conducted while implementing new wage equations in the Insee microsimulation model Destinie 2, see Cazenave-Lacroutz *et al.* (2019). Alexandre thanks Prof.Dr. Winter for having hosted him at LMU University of Munich when this research started.

[a] Université Paris Dauphine, Place du Maréchal de Lattre de Tassigny, 75016 Paris

[b] Crest, 5 Avenue Henry Le Chatelier, 91120 Palaiseau

1 Introduction

This note considers the estimation of linear unobserved effects panel data models[1] with AR(1) disturbances, that refer to processes of the form:

$$y_{it} = x'_{it}\beta + \nu_i + u_{it} \tag{1}$$

$$u_{it} = \rho u_{it-1} + \varepsilon_{it} \tag{2}$$

where $|\rho| < 1$ and the ε_{it}'s are i.i.d. disturbances with mean 0 and variance σ_ε^2.

Such estimations are common in the wage equation literature (with even higher order of correlation), and are not infrequent in the general economics literature. They are performed in Stata with the command *xtregar*, which has been used in influential and recent economics articles such as Dafny (2010), Hau *et al.* (2013) or Prieto et Lago-Peñas (2012).

In Cazenave-Lacroutz et Lin (2019), we proposed new estimators for the auto-correlation coefficient. In the particular case of random-effects setting (that is: ν_i is exogenous to the covariates) and a positive auto-correlation parameter[2], it enables to produce consistent estimators for all parameters of the model, assuming there are enough individuals with at least two consecutive observations and at least three observations overall.

We now turn to the more general case when no hypothesis is made regarding the exogeneity of the unobserved individual effect. First we remind the properties of the current estimation method implemented in the Stata command **xtregar, fe**. Additionnaly, we explain why it works very well to estimate the variance of the perturbation in the balanced case, but fails to do so in the unbalanced case. We also suggest a novel estimation method for the coefficients of the independent variables, as the current estimation method may perform poorly in some circumstances. Second, a consistent estimator of the variance of the perturbations is derived. Third, we illustrate the properties of these estimators though Monte Carlo simulations that

[1] That is in both fixed and random effects models.

[2] In case of a negative auto-correlation parameter, asymptotically consistent estimator is provided, where the convergence is achieved when the minimum number of period per individual goes to infinity.

compare the respective performances of the (more specific) random-effects estimators and of the (more general) fixed-effects estimators. We eventually conclude.

2 The current estimation method

In all the paper, we consider that a consistent estimator of the autocorrelation parameter ρ is known (Cazenave-Lacroutz et Lin, 2019). To alleviate the notations, we even consider that the true value of ρ is known.

When the fixed-effects are assumed to be exogeneous to the covariates (the random-effects model), Baltagi et Wu (1999) propose a transformation of the data that enables to get consistent estimates of the β parameter, of the variance of the fixed-effects, and of the variance of the perturbations.

In the general case (when there is no hypothesis regarding the exogeneity of the fixed-effects), the current estimation method (for instance implemented by the Stata command xtregar) first applies to the fixed-effects model the Baltagi-and-Wu transformation (Baltagi et Wu, 1999). We detail it below.

2.1 The current estimation method

Baltagi et Wu (1999) derive a transformation $C_i(\rho)$ of the data that removes the AR(1) component. Following this transformation, one gets:

$$\begin{aligned} y^*_{i,t_{i,j}} &= (1-\rho^2)^{1/2} y_{i,t_{i,j}} \text{ if } j == 1 \\ &= (1-\rho^2)^{1/2}\left(y_{i,t_{i,j}} \frac{1}{(1-\rho^{2(t_{i,j}-t_{i,j-1})})^{1/2}} - y_{i,t_{i,j-1}} \frac{\rho^{t_{i,j}-t_{i,j-1}}}{(1-\rho^{2(t_{i,j}-t_{i,j-1})})^{1/2}}\right) \text{ if } j > 1 \end{aligned} \tag{3}$$

Let us consider equation (1):

$$y_{it} = x'_{it}\beta + \nu_i + u_{it}$$

By applying the Baltagi-and-Wu transformation, one gets:

$$y^*_{it} = x^{*'}_{it}\beta + \nu^*_{i,t} + u^*_{it} \tag{4}$$

Quite importantly, note that the fixed effects ν_i have became $\nu^*_{i,t}$. In the general case, the Baltagi-transformed fixed effects are no longer fixed over time.[3]

It is easy to show that the error terms u^*_{it} are no longer correlated and are homoskedastic.

The current estimation method (see Stata - xtregar / Methods and Formula) then differentes this equation with the mean and grand mean. Let us note for a variable x, with $n_i = \sum_{t=1}^{T} 1(]j_0/t_{i,j_0} == t)$:

$$\overline{x^*} = \frac{\sum_{j=2}^{n_i} x^*_{i,t(i,j)}}{n_i - 1}$$

$$\overline{\overline{x^*}} = \frac{\sum_{i=1}^{N}\sum_{j=2}^{n_i} x^*_{i,t(i,j)}}{\sum_{i=1}^{N}(n_i - 1)}$$

$$x^{**}_{it} = x^*_{it} - \overline{x^*} + \overline{\overline{x^*}} \tag{5}$$

The transformed equation is thus:

$$y^{**}_{it} = x^{**'}_{it}\beta + \nu^{**}_{it} + u^{**}_{it} \tag{6}$$

An OLS regression of y^{**} on the x^{**} and a constant is then performed.

2.2 The estimation of the β

It is trivial to see that the current methods enables to get consistent estimates of the β in the balanced case, as the $\nu_{i,t}$ are independent of t in this very particular case and are thus dropped due to the demeaning. This note makes no contribution to that regard.

In the unbalanced case, even if this method seems to provide consistent estimators of the β in many of our Monte-Carlo simulations, we are able to show its validity only under specific hypotheses (see Annex A). In Section 5, we also present a Monte

[3] The balanced case is an exception to that regard, as in this very particular case: $v^*_i = (1 - \rho)\nu_i$

Carlo simulation where β does not seem to be consistently estimated.

Happily, we are able to propose a novel (consistent) estimation of the β in Section 3.

In addition, note that there is no reason in the above methods that the identified constant respects the usual convention[4] that the sum of the ν_i is equal to zero.

2.3 An imprecise estimation of the fixed effects ν_i

Once a consistent estimator of β has been found, the parameters ν_i can be estimated by:

$$\hat{\nu_i} = \bar{y_{i,t}} - (\bar{x_{i,t}}'\hat{\beta}) \tag{7}$$

There might be made centered around zero. Even without centering it, the variance of these estimates provides an estimate of the variance of the fixed-effects. As such variance is based on the imprecise estimates of the individual fixed-effects, it is quite imprecise. It converges however towards the true variance when the minimal number of observation per individual tends to infinity.

2.4 The variance of the perturbations

We focus here on the current estimation of the variance of the perturbations σ_{ε}^2. One estimation method in the balanced case it to take as an estimator of σ_{ε}^2 the empirical variance of $u_{i,t}^*$, which yields a consistent estimator of σ_{ε}^2. But in the unbalanced case, there is no reason why it would yield a consistent estimator of σ_{ε}^2. Monte Carlo simulations in Section 5 clearly shows that the estimates of the current command **xtregar** can be far away from the true value of σ_{ε}^2.

[4] This convention is usual in Stata.

<u>The balanced case:</u>

By applying the Baltagi-and-Wu transformation to the balanced case to the perturbation $u_{i,t}$, it comes:

$$\begin{aligned} u^*_{i,t} &= (1-\rho^2)^{1/2} \text{ if } t == 1 \\ &= (1-\rho^2)^{1/2}(u_{i,t}\frac{1}{(1-\rho^2)^{1/2}} - u_{i,t-1}\frac{\rho}{(1-\rho^2)^{1/2}}) \text{ if } t > 1 \end{aligned}$$

We remind equation (2):

$$u_{i,t} = \rho u_{i,t-1} + \varepsilon_{it}$$

This enables to conclude as we easily get that:

$$u^*_{i,t} = \varepsilon_{it}$$

$$\begin{aligned} \text{Var}(u^{**}_{it}) &= (1 + \frac{N-1}{N(T-1)}(\frac{2}{T-1} + \frac{1}{N} - 3))\sigma^2_\varepsilon \\ &\neq \sigma^2_\varepsilon \end{aligned}$$

3 Consistent estimators of β

A first method to get a consistent estimator of β consists in estimating the model without taking into account the autocorrelation of the perturbations. That is: one estimates equation (1) but not taking into account equation (2). One needs to adjust the computation of the standard errors of the estimated coefficients to the fact that there is some intra-individual autocorrelation.[5]

As a second (natural) method of estimation of the β, we propose a modification of the Baltagi-and-Wu transformation that yields another consistent estimator of the β coefficient.It delivers a more efficient estimator of the β than with the first method of estimation above in some circumstances. Note that in the balanced case, it yields exactly the same estimator of β as if the Baltagi-and-Wu tranformation had been applied.

[5] In practice, in Stata, one should use: **xtreg, fe vce(cluster id)**.

We observe that:

$$y^*_{it} = x^{*'}_{it}\beta + (1-\rho^2)^{1/2}1_{i,t}\nu_i + u^*_{it}$$

Where:

$$\begin{aligned} 1_{i,t_{i,j}} &= 1 \text{ if } j == 1 \\ &= \frac{1-\rho^{t_{i,j}-t_{i,j-1}}}{(1-\rho^{2(t_{i,j}-t_{i,j-1})})^{1/2}} \text{ if } j > 1 \end{aligned} \tag{8}$$

We propose thus the following transformation that make the transformed fixed-effects to remain fixed over time, and the transformed perturbation to be uncorrelated:

$$x^{*2}_{it} = \frac{1}{1_{i,t}}x^*_{it} \tag{9}$$

Hence, our modification of the Baltagi-and-Wu transformation is the following:

$$\begin{aligned} x^{*2}_{i,t_{i,j}} &= (1-\rho^2)^{1/2}x_{i,t_{i,j}} \text{ if } j == 1 \\ &= (1-\rho^2)^{1/2}(x_{i,t_{i,j}}\frac{1}{(1-\rho^{(t_{i,j}-t_{i,j-1})})} - x_{i,t_{i,j-1}}\frac{\rho^{t_{i,j}-t_{i,j-1}}}{(1-\rho^{(t_{i,j}-t_{i,j-1})})}) \text{ if } j > 1 \end{aligned} \tag{10}$$

From that point on, the principle is the following: by demeaning as in the current method, we get rid of all the terms with ν_i:

$$y^{*2}_{it} - \bar{y}^{*2}_{it} = (x^{*2'}_{it} - \bar{x}^{*2'}_{it})\beta + (u^{*2}_{it} - \bar{u}^{*2}_{it})$$

An OLS estimation of the above regression yields a consistent estimate of β.

In practice, the demeaning is performed using Stata **xtreg, fe vce(robust)** command, as it handles the remaining heteroskedasticity and serial correlation in the transformed errors u^{*2}_{it}. The above Stata command indeed takes into account the note of Stock et Watson (2008) that tackles both issues.[6]

[6] Note that the variance-covariance matrix is a diagonal matrix whose coefficients are $\frac{\sigma_\epsilon}{1^2_{i,t}}$. As it is explicitly known, a better estimator of the variance of the β that takes into account this explicit form is possible. This is let to further work.

A comparison of the two methods

None of the above method is necessarily more efficient than the other. Which one is more efficient depends on the pattern of the missing observations, and the relative importance of σ_ϵ and σ_ν.

The first method identifies β based on the following equation, to which demeaning is then applied:

$$y_{it} = x'_{it}\beta + \nu_i + u_{it}$$

The second method identifies β based on the following equation, to which demeaning is then applied:

$$y^{*2}_{it} = x^{*2'}_{it}\beta + \nu^{*2}_i + u^{*2}_{it}$$

Both equations are fixed-effect equations whose terms are uncorrelated. There are however homoscedastic of variance σ_ϵ in the first above equation ; and heteroscedastic of variance $\frac{\sigma_\eta}{1^2_{i,t}}$.

To begin with, we consider the balanced case. The terme $1^2_{i,t}$ is a constant and can be discarded as it multiplies all the terms of the regressions. As $\sigma_\eta = (1-\rho^2)\sigma_\epsilon$, the variance of the perturbations in the second equation σ_η is lower than the variance of the perturbations in the first equation σ_ϵ. This makes the second estimator of β more efficient than the estimator of β that does not take into account the autocorrelations of the perturbations.

However, this is likely to change in the unbalanced case when $1^2_{i,t}$ is no longer a constant. We provide Monte Carlo simulations where the first method of estimation is more efficient than the second method of estimation (e.g. Table 1). We expect that our second method of estimation of the β is more efficient than the first method the closer to the balanced case. Except in the balanced case where the Baltagi-and-Wu transformation is unambiguously more efficient, we suggest to the interested researchers to perform the two analyses and to choose the one whose reported standard errors of β are the lower.

4 A new estimator of the variance of the perturbations

A naive estimator of σ_ε^2 can be obtained by explicitly computing $\hat{u}_{i,t} := y_{i,t} - x'_{i,t}\hat{\beta} - \hat{\nu}_i$; considering its empirical variance, and multiplying it by $1 - \rho^2$ to obtain an estimator σ_ε^2. It however yields an imprecise estimator of σ_ε, as it relies on the imprecise estimation of the ν_i. We therefore propose an estimator that is not based on the estimation of the ν_i.

To do so, we define:

$$\widetilde{y}_{i,t} = y_{i,t} - x'_{i,t}\beta \tag{11}$$

As highlighted above, under the missing-at-random hypothesis, the current estimation method enables to get a consistent estimate of β, and thus to compute a consistent estimate of the above quantity.
We observe that:

$$\widetilde{y}_{i,t} = y_{i,t} - x'_{i,t}\beta = \nu_i + u_{i,t} \tag{12}$$

We define :

$$\widetilde{\widetilde{y}}_{i,t(i,j)} = \widetilde{y}_{i,t(i,j)} - \widetilde{y}_{i,t(i,j-1)} = u_{i,t(i,j)} - u_{i,t(i,j-1)} \tag{13}$$

By successively applying equation (2), it comes:

$$\widetilde{\widetilde{y}}_{i,t(i,j)} = (\rho^{t(i,j)-t(i,j-1)} - 1)u_{i,t(i,j-1)} + \sum_{k=0}^{t(i,j)-t(i,j-1)-1} \rho^k \varepsilon_{i,t(i,j)-k} \tag{14}$$

Hence, since all the above terms in the sum are uncorrelated and of mean 0:

$$\mathrm{E}(\widetilde{\widetilde{y}}\,^2_{i,t(i,j)}) = \mathrm{Var}(\widetilde{\widetilde{y}}_{i,t(i,j)}) = (1 - \rho^{t(i,j)-t(i,j-1)})^2\sigma_u^2 + \sum_{k=0}^{t(i,j)-t(i,j-1)-1} \rho^{2k}\sigma_\varepsilon^2$$

Thus:

$$\sigma_\varepsilon^2 = \frac{E(\widetilde{\widetilde{y}}\,^2_{i,t(i,j)})}{\frac{(1-\rho^{t(i,j)-t(i,j-1)})^2}{(1-\rho^2)} + \frac{(1-\rho^{2(t(i,j)-t(i,j-1))})}{(1-\rho^2)}} \tag{15}$$

We have obtained:

$$\sigma_\varepsilon^2 = \mathrm{E}(w_{i,t}) \tag{16}$$

where:

$$w_{i,t} = \frac{\widetilde{\widetilde{y}}\,{}^{2}_{i,t(i,j)}}{\frac{(1-\rho^{t(i,j)-t(i,j-1)})^2}{(1-\rho^2)} + \frac{(1-\rho^{2(t(i,j)-t(i,j-1))})}{(1-\rho^2)}} \tag{17}$$

Under very general hypotheses, one gets a natural consistent estimate of σ_ε^2 (see Annex B for a demonstration):

$$\widehat{\sigma_\varepsilon^2} = \frac{1}{N}\sum_{i=1}^{N} \bar{w}_{i,t} \tag{18}$$

5 Monte Carlo simulations

To illustrate the previous theoretical sections, we perform Monte Carlo simulations with the following basis parameters : $N = 500$, $T = 10$, $\rho = 0.6$, $\sigma_\varepsilon = 0.3$, $\sigma_\nu = 0.35$. Those parameters were already used by Cazenave-Lacroutz et Lin (2019), enabling comparability [7].

5.1 In the random effects design

First, we consider the case where the individual effects ν_i are exogeneous from the covariables x[8]. This is a very particular case, where a random-effects estimation method can also be applied. This allows to compare the estimation advantages of making the assumption of a random-effects model (when suitable) rather than the more general fixed-effects model.

With both random-effects and fixed-effects models, we note that the estimation of β is (slightly) more efficient with **xtreg** than the one we proposed. In the general fixed-effects case, we also note that in this particular case, the estimation of β by xtregar does not seem biased, although it is far less efficient than our method of estimation or the one of **xtreg**.

[7] The values of ρ, σ_ε and σ_ν were chosen by Cazenave-Lacroutz et Lin (2019), as they were typical of what they encountered in an applied study, see Cazenave-Lacroutz *et al.* (2019). The values $T = 10$ and $N = 500$ ensure that the panel considered is short in the time dimension but with an number of individual units close to infinity.

[8] Typically, we consider covariables as a random draw that has an additive impact on the dependent variable, and a constant.

With our estimation method for σ_ϵ, both the fixed-effects method and the random-effects method provide consistent estimates of the standard deviation of the perturbations (see Table 1). However, due to the limited number of periods observed per individual, the estimation of the standard deviation σ_ν derived from the fixed effects method is biased, but unbiased with the random-effect method[9].

We also note that in this particular case, the coefficient of the covariable is consistently estimated by both estimation methods.

Table 1. Monte Carlo simulations on an unbalanced panel, $T = 10$

	Fixed-effects method				Random-effects method			
	ρ	σ_ε	σ_ν	*covar*	ρ	σ_ε	σ_ν	*covar*
true values	0.6	0.3	0. 35	3	0.6	0.3	0. 35	3
with true ρ and xtregar	.6 (0)	.621*** (4.9e-03)	.447*** (.013)	3 (.01)	.6 (0)	.3 (4.2e-03)	.356 (.016)	3 (6.4e-03)
with true ρ and corrected xtregar	.6 (0)	.3 (4.6e-03)	.447*** (.013)	3 (6.7e-03)	.6 (0)	.3 (4.2e-03)	.356 (.016)	3 (6.4e-03)
with ρ_{BFN} and corrected xtregar	.581 (.04)	.299 (3.2e-03)	.447*** (.013)	3 (6.5e-03)	.581 (.04)	.3 (3.9e-03)	.359 (.023)	3 (6.3e-03)
with true ρ and xtreg	0 (0)	.323*** (5.9e-03)	.432*** (.013)	3 (5.2e-03)	0 (0)	.323*** (5.9e-03)	.403*** (.014)	3 (5.1e-03)

Legend: The average estimators should not be significantly different from the true values. It is the case only for those in bold. Significance levels for the differences with the true values are otherwise pinpointed by stars: * $(p < 0.10)$, ** $(p < 0.05)$, *** $(p < 0.01)$

Note 1: Approximately half of a panel of 500 individuals observed each over T = 10 periods has been randomly deleted, before the Monte Carlo process has been implemented with 50 replications.

The estimates of σ_ε and of σ_ν in the two last lines are obtained by estimating first ρ_{BFN} (or ρ_{BFN2U}), and then by imposing it as the estimate of ρ in *xtregar*.

In Table 2, the same simulations are performed with a longer time period ($T = 100$

[9] Indeed, in the random-effects method, it can be estimated without relying on the imprecise estimation of the various ν_i.

rather than $T = 10$). Accordingly with our previous analysis, the fixed-effects method yields estimates of the standard deviation σ_ν that are no longer significantly different from its true value. Unlike what was observed in Table 1, With both random-effects and fixed-effects models, we also note that the estimation of β is (slightly) more efficient with our method of estimation than with **xtreg**.

Table 2. Monte Carlo simulations on an unbalanced panel, $T = 100$

	Fixed-effects method				Random-effects method			
	ρ	σ_ε	σ_ν	*covar*	ρ	σ_ε	σ_ν	*covar*
true values	0.6	0.3	0. 35	3	0.6	0.3	0. 35	3
with true ρ and xtregar	.6	.678***	.36	3	.6	.3	.35	3
	(0)	(2.8e-03)	(.01)	(4.9e-03)	(0)	(1.1e-03)	(9.9e-03)	(2.4e-03)
with true ρ and corrected xtregar	.6	.3	.36	3	.6	.3	.35	3
	(0)	(1.3e-03)	(.01)	(2.7e-03)	(0)	(1.1e-03)	(9.9e-03)	(2.4e-03)
with ρ_{BFN} and corrected xtregar	.601	.3	.36	3	.601	.3	.35	3
	(4.6e-03)	(1.2e-03)	(.01)	(2.7e-03)	(4.6e-03)	(1.2e-03)	(9.9e-03)	(2.4e-03)
with xtreg	0	.37***	.36	3	0	.37***	.356	3
	(0)	(1.8e-03)	(1.0e-02)	(3.0e-03)	(0)	(1.8e-03)	(.01)	(3.0e-03)

Legend: The average estimators should not be significantly different from the true values. It is the case only for those in bold. Significance levels for the differences with the true values are otherwise pinpointed by stars: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$)

Note 1: Approximately half of a panel of 500 individuals observed each over T = 100 periods has been randomly deleted, before the Monte Carlo process has been implemented with 50 replications.

The estimates of σ_ε and of σ_ν in the two last lines are obtained by estimating first ρ_{BFN} (or ρ_{BFN2U}), and then by imposing it as the estimate of ρ in *xtregar*.

5.2 In the general case

We introduce two changes in regard with the simulations presented in Table 1.
First, the individual effects are no longer exogeneous to the covariates[10]. Hence, we no longer present the results from the random-effects method, as it is based on this assumption of orthogonality.
Second, while the data were missing at random, we deviate from this missing pattern. In the "Non missing at random" setting, missingness is based on the value taken by the covariates. As shown in Table 3, this yields an estimate for the coefficient of the covariable that is significantly different from its true value, which was not the case under the Missing-at-random setting. Until now, we do not know to which extent the absence of the Missing-at-random hypothesis is central in explaining this. Our main purpose is to show that the parameter β is not always consistently estimated under the current procedure. As expected, our corrected procedure yields a consistent estimate in both studied cases.

[10] More precisely, the covariate is the sum of a random term and the fixed-effect.

Table 3. Monte Carlo simulations on an unbalanced panel, $T = 10$, general case

	Missing-at-random				Non missing-at-random			
	ρ	σ_ε	σ_ν	covar	ρ	σ_ε	σ_ν	covar
true values	0.6	0.3	0. 35	3	0.6	0.3	0. 35	3
with true ρ and xtregar	.6	.621***	.442***	3	.6	.458***	.377	2.95***
	(0)	(4.9e-03)	(.019)	(.012)	(0)	(.017)	(.021)	(.013)
with true ρ and corrected xtregar	.6	.3	.447***	3	.6	.302	.355	3
	(0)	(4.6e-03)	(.016)	(6.7e-03)	(0)	(7.5e-03)	(.018)	(8.0e-03)
with ρ_{BFN} and corrected xtregar	.581	.299	.447***	3	.581	.3	.355	3
	(.04)	(3.2e-03)	(.016)	(6.5e-03)	(.04)	(8.3e-03)	(.018)	(8.0e-03)
with xtreg	0	.323***	.434***	3	0	.326***	.373	2.99
	(0)	(5.9e-03)	(.015)	(5.2e-03)	(0)	(7.0e-03)	(.021)	(.013)

Legend: The average estimators should not be significantly different from the true values. It is the case only for those in bold. Significance levels for the differences with the true values are otherwise pinpointed by stars: * ($p < 0.10$), ** ($p < 0.05$), *** ($p < 0.01$)

Note 1: Approximately half of a panel of 500 individuals observed each over T = 10 periods has been randomly deleted, before the Monte Carlo process has been implemented with 50 replications.

The estimates of σ_ε and of σ_ν in the two last lines are obtained by estimating first ρ_{BFN} (or ρ_{BFN2U}), and then by imposing it as the estimate of ρ in *xtregar*.

Note 2: *covar* is an independent variable that is the sum of a random term and the fixed-effect.

6 Conclusion

Whereas Baltagi et Wu (1999) proposed a way of consistently estimating the parameters of a random-effects regression with AR(1) perturbations[11], to the best of our knowledge, no previous paper attempted to generalize their method to the case where no hypothesis is made regarding the exogeneity of the fixed effects. Current practice consisted in applying the Baltagi-and-Wu transformation before demeaning. We show that this procedure is adapted in the very particular case of balanced panels, but not necessarily in the more common case of unbalanced panel.

In the unbalanced case, we provide Monte Carlo simulations where the current procedure does not yield a consistent estimation of the β parameter. Furthermore, we propose a novel transformation that provides consistent estimates also in the unbalanced case (that may be more efficient than estimating β by ignoring the autocorrelation of the perturbations). We also notice that the constant does not *a priori* respects the nullity convention of the sum of the fixed effects, but the constant is usually not an object of interest *per se.* More importantly for some applications (e.g. simulations generated by estimated parameters), the variance of the perturbation was not consistently estimated. In case a consistent estimate of β is available, we propose an additional estimation procedure that enables to get a consistent estimator of σ_ϵ. Hence, all parameters of this type of model are thus consistently estimated[12].

References

Baltagi, B. H. et Wu, P. X. (1999). Unequally spaced panel data regressions with AR (1) disturbances. *Econometric Theory*, 15(6):814–823.

Cazenave-Lacroutz, A., Godet, F. et Lin, V. (2019). Modélisation des trajectoires de revenus d'activité pour le modèle destinie 2.

Cazenave-Lacroutz, A. et Lin, V. (2019). The estimation of the autocorrelation coefficient in panel data models with ar(1) disturbances.

[11] when a consistent estimator of the autocorrelation parameter ρ is available. Such a consistent estimator is proposed by Cazenave-Lacroutz et Lin (2019).

[12] The variance of the fixed-effects is consistently estimated as T tends to ∞.

Dafny, L. S. (2010). Are health insurance markets competitive? *American Economic Review*, 100(4):1399–1431.

Hau, H., Langfield, S. et Marques-Ibanez, D. (2013). Bank ratings: what determines their quality? *Economic Policy*, 28(74):289–333.

Prieto, D. C. et Lago-Peñas, S. (2012). Decomposing the determinants of health care expenditure: the case of spain. *The European Journal of Health Economics*, 13(1):19–27.

Stock, J. H. et Watson, M. W. (2008). Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica*, 76(1):155–174.

Annexes

A Consistency of $\hat{\beta}$

We focus on the consistency of $\hat{\beta}$, where $\hat{\beta}$ denotes the OLS estimator of β obtained from equation (6). It suffices to show that the orthogonality condition $\mathbb{E}(x^{**}_{it_{ij}}(\nu^{**}_{it_{ij}} + u^{**}_{it_{ij}})) = 0$ holds. First, the relation $\mathbb{E}(x^{**}_{it_{ij}}u^{**}_{it_{ij}}) = 0$ trivially follows from the strict exogeneity assumptions $\mathbb{E}(x_{it_{ij}}u_{kt_{kl}}) = 0$. According to Lemma 1, the second relation $\mathbb{E}(x^{**}_{it_{ij}}\nu^{**}_{it_{ij}}) = 0$ holds if one of two conditions holds.

Lemma 1: Condition $\mathbb{E}(x^{**}_{it_{ij}}\nu^{**}_{it_{ij}}) = 0$ holds for all $2 \leq j \leq n_i$
if:

$$\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}} = (1 - \frac{n_i - 1}{\sum\limits_{i=1}^{N}(n_i-1)})\frac{1}{n_i-1}\sum_{l=2}^{n_i}\frac{1-\rho^{t_{il}-t_{il-1}}}{\sqrt{1-\rho^{2(t_{il}-t_{il-1})}}} \tag{19}$$

for all $2 \leq j \leq n_i$.

or if:

$$\mathbb{E}(x^{**}_{it_{ij}}\nu_i) = 0 \tag{20}$$

First, we compute

$$\begin{aligned}\mathbb{E}(x^{**}_{it_{ij}}\nu^{*}_{it_{ij}}) &= \mathbb{E}(x^{**}_{it_{ij}}\sqrt{1-\rho^2}\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}\nu_i) \\ &= \sqrt{1-\rho^2}\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}\mathbb{E}(x^{**}_{it_{ij}}\nu_i),\end{aligned}$$

$$\begin{aligned}\mathbb{E}(x^{**}_{it_{ij}}\overline{\nu^*_i}) &= \mathbb{E}(\frac{1}{n_i-1}\sum_{l=2}^{n_i} x^{**}_{it_{ij}}\nu^*_{it_{il}}) \\ &= \mathbb{E}(\frac{1}{n_i-1}\sum_{l=2}^{n_i}\sqrt{1-\rho^2}\frac{1-\rho^{t_{il}-t_{il-1}}}{\sqrt{1-\rho^{2(t_{il}-t_{il-1})}}}x^{**}_{it_{ij}}\nu_i) \\ &= \frac{\sqrt{1-\rho^2}}{n_i-1}\sum_{l=2}^{n_i}\frac{1-\rho^{t_{il}-t_{il-1}}}{\sqrt{1-\rho^{2(t_{il}-t_{il-1})}}}\mathbb{E}(x^{**}_{it_{ij}}\nu_i)\end{aligned}$$

$$\begin{aligned}\mathbb{E}(x^{**}_{it_{ij}}\overline{\overline{\nu^*}}) &= \frac{1}{\sum\limits_{l=1}^{N}(n_l-1)}\mathbb{E}(\sum_{l=1}^{N}\sum_{k=2}^{n_l}x^{**}_{it_{ij}}\nu^*_{lt_{lk}}) \\ &= \frac{1}{\sum\limits_{l=1}^{N}(n_l-1)}\mathbb{E}(\sum_{k=2}^{n_i}x^{**}_{it_{ij}}\nu^*_{it_{ik}}) \\ &= \frac{n_i-1}{\sum\limits_{l=1}^{N}(n_l-1)}\mathbb{E}(x^{**}_{it_{ij}}\overline{\nu^*_i}),\end{aligned}$$

the second equality following from the independency between $x^{**}_{it_{ij}}$ and $\nu^*_{lt_{lk}}$ for $l \neq i$, which yields $\mathbb{E}(x^{**}_{it_{ij}}\nu^*_{lt_{lk}}) = \mathbb{E}(x^{**}_{it_{ij}})\mathbb{E}(\nu^*_{lt_{lk}}) = 0$.

Since $\nu^{**}_{it_{ij}} = \nu^*_{it_{ij}} - \overline{\nu^*_i} + \overline{\overline{\nu^*}}$, we have:

$$\begin{aligned}&\mathbb{E}(x^{**}_{it_{ij}}\nu^{**}_{it_{ij}}) = \mathbb{E}(x^{**}_{it_{ij}}(\nu^*_{it_{ij}} - \overline{\nu^*_i} + \overline{\overline{\nu^*}})) \\ &= \sqrt{1-\rho^2}\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}\mathbb{E}(x^{**}_{it_{ij}}\nu_i) - (1-\frac{n_i-1}{\sum\limits_{i=1}^{N}(n_i-1)})\mathbb{E}(x^{**}_{it_{ij}}\overline{\nu^*_i}) \\ &= \sqrt{1-\rho^2}\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}\mathbb{E}(x^{**}_{it_{ij}}\nu_i) - (1-\frac{n_i-1}{\sum\limits_{i=1}^{N}(n_i-1)})\frac{\sqrt{1-\rho^2}}{n_i-1}\sum_{l=2}^{n_i}\frac{1-\rho^{t_{il}-t_{il-1}}}{\sqrt{1-\rho^{2(t_{il}-t_{il-1})}}}\mathbb{E}(x^{**}_{it_{ij}}\nu_i)\end{aligned}$$

Hence $\mathbb{E}(x^{**}_{it_{ij}}\nu^{**}_{it_{ij}}) = 0$ if and only if condition (19) holds or $\mathbb{E}(x^{**}_{it_{ij}}\nu_i) = 0$. This demonstrates Lemma 1.

We first consider equation (19). It writes:

$$\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}} = (1 - \frac{n_i - 1}{\sum\limits_{i=1}^{N}(n_i - 1)})\frac{1}{n_i - 1}\sum_{l=2}^{n_i}\frac{1-\rho^{t_{il}-t_{il-1}}}{\sqrt{1-\rho^{2(t_{il}-t_{il-1})}}}$$

for all $2 \leq j \leq n_i$.

If these conditions were true, the quantities $\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}$, $2 \leq j \leq n_i$, would all be equal.
In the very particular case of the balanced case, all these conditions would be summed up in one single condition:

$$1 = 1 - \frac{1}{N}$$

It approximately holds for large N.
Hence, if we are in a balanced panel, these conditions are approximately respected.

Note that in the unbalanced case, if the data pattern were random, then condition (19) would be valid in expectancy (and for N large enough).

Lemma 2: Under the hypothesis that missing occurs independently of the observable $x_{i,t}$ and of the fixed-effects ν_i (i.e. the *missing-at-random* hypothesis), equation (19) holds in expectancy for N large enough.

We compute

$$\begin{aligned}
&\mathbb{E}((1-\frac{n_i-1}{\sum\limits_{i=1}^{N}(n_i-1)})\frac{1}{n_i-1}\sum_{l=2}^{n_i}\frac{1-\rho^{t_{il}-t_{il-1}}}{\sqrt{1-\rho^{2(t_{il}-t_{il-1})}}}|n_1,...,n_N)\\
&=(1-\frac{n_i-1}{\sum\limits_{i=1}^{N}(n_i-1)})\frac{1}{n_i-1}\sum_{l=2}^{n_i}\mathbb{E}(\frac{1-\rho^{t_{il}-t_{il-1}}}{\sqrt{1-\rho^{2(t_{il}-t_{il-1})}}}|n_1,...,n_N)\\
&=(1-\frac{n_i-1}{\sum\limits_{i=1}^{N}(n_i-1)})\frac{1}{n_i-1}(n_i-1)\mathbb{E}(\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}|n_1,...,n_N)\\
&=(1-\frac{n_i-1}{\sum\limits_{i=1}^{N}(n_i-1)})\mathbb{E}(\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}|n_1,...,n_N)\\
&=\mathbb{E}((1-\frac{n_i-1}{\sum\limits_{i=1}^{N}(n_i-1)})\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}|n_1,...,n_N)
\end{aligned}$$

the second equality following from the fair assumption that $\mathbb{E}(\frac{1-\rho^{t_{il}-t_{il-1}}}{\sqrt{1-\rho^{2(t_{il}-t_{il-1})}}}|n_i)$ does not depend on l. Taking expectations, we have

$$\begin{aligned}
&\mathbb{E}((1-\frac{n_i-1}{\sum\limits_{i=1}^{N}(n_i-1)})\frac{1}{n_i-1}\sum_{l=2}^{n_i}\frac{1-\rho^{t_{il}-t_{il-1}}}{\sqrt{1-\rho^{2(t_{il}-t_{il-1})}}})\\
&=\mathbb{E}((1-\frac{n_i-1}{\sum\limits_{i=1}^{N}(n_i-1)})\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}})
\end{aligned}$$

Then, condition 19 holds in expectancy if and only if

$$\mathbb{E}(\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}})=\mathbb{E}((1-\frac{n_i-1}{\sum\limits_{i=1}^{N}(n_i-1)})\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}})$$

which is the case when N grows to infinity.

We focus here on the condition $\mathbb{E}(x^{**}_{it_{ij}}\nu_i) = 0$.
First, we compute

$$\mathbb{E}(x^*_{it_{ij}}\nu_i) = \frac{\sqrt{1-\rho^2}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}\mathbb{E}((x_{it_{ij}} - \rho^{t_{il}-t_{il-1}}x_{it_{ij-1}})\nu_i)$$

$$\mathbb{E}(\overline{x^*_i}\nu_i) = \frac{\sqrt{1-\rho^2}}{n_i-1}\sum_{j=2}^{n_i}\mathbb{E}(\frac{x_{it_{ij}} - \rho^{t_{il}-t_{il-1}}x_{it_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}\nu_i)$$

$$\mathbb{E}(\overline{\overline{x^*}}\nu_i) = \frac{n_i-1}{\sum\limits_{l=1}^{N}(n_l-1)}\mathbb{E}(\overline{x^*_i}\nu_i)$$

Hence

$$\begin{aligned}
&\mathbb{E}(x^{**}_{it_{ij}}\nu_i) = \mathbb{E}((x^*_{it_{ij}} - \overline{x^*_i} + \overline{\overline{x^*}})\nu_i) \\
&= \frac{\sqrt{1-\rho^2}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}\mathbb{E}((x_{it_{ij}} - \rho^{t_{il}-t_{il-1}}x_{it_{ij-1}})\nu_i) - (1 - \frac{n_i-1}{\sum\limits_{l=1}^{N}(n_l-1)})\mathbb{E}(\overline{x^*_i}\nu_i) \\
&= \frac{\sqrt{1-\rho^2}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}\mathbb{E}((x_{it_{ij}} - \rho^{t_{il}-t_{il-1}}x_{it_{ij-1}})\nu_i) \\
&- (1 - \frac{n_i-1}{\sum\limits_{l=1}^{N}(n_l-1)})\frac{\sqrt{1-\rho^2}}{n_i-1}\sum_{j=2}^{n_i}\mathbb{E}(\frac{x_{it_{ij}} - \rho^{t_{il}-t_{il-1}}x_{it_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}}\nu_i)
\end{aligned}$$

A sufficient condition for $\mathbb{E}(x^{**}_{it_{ij}}\nu_i) = 0$ is that $\mathbb{E}(x^*_{it_{ij}}\nu_i)$ does not depend on j. Indeed, in this case,

$$\mathbb{E}(x^{**}_{it_{ij}}\nu_i) = \frac{n_i-1}{\sum\limits_{l=1}^{N}(n_l-1)}\mathbb{E}(x^*_{it_{ij}}\nu_i),$$

which converges to 0 when N tends to infinity.

Note however that the necessary hypothesis for this result $((x^*_{it_{ij}}\nu_i)$ does not depend on $j)$ depends on the value of ρ (through the Baltagi-and-Wu transformation) and is thus not necessarily very general.

B Proof of consistency of $\hat{\sigma_\epsilon}$

We have defined:

$$\widehat{\sigma_\varepsilon^2} = \frac{1}{N}\sum_{i=1}^{N} \bar{w}_{i,t}$$

where:

$$\mathrm{E}(w_{i,t}) = \sigma_\varepsilon^2$$

and with:

$$w_{i,t} = \frac{\widetilde{\widetilde{y}}\,^{2}_{i,t(i,j)}}{\frac{(1-\rho^{t(i,j)-t(i,j-1)})^2}{(1-\rho^2)} + \frac{(1-\rho^{2(t(i,j)-t(i,j-1))})}{(1-\rho^2)}}$$

That is:

$$w_{i,t} = \frac{\left(u_{i,t(i,j)} - u_{i,t(i,j-1)}\right)^2}{\frac{(1-\rho^{t(i,j)-t(i,j-1)})^2}{(1-\rho^2)} + \frac{(1-\rho^{2(t(i,j)-t(i,j-1))})}{(1-\rho^2)}}$$

$$w_{i,t} = \frac{\left(\sum_{k=1}^{t(i,j)-t(i,j-1)} \rho^{k-1}\eta_{t(i,j)-k}\right)^2}{\frac{(1-\rho^{t(i,j)-t(i,j-1)})^2}{(1-\rho^2)} + \frac{(1-\rho^{2(t(i,j)-t(i,j-1))})}{(1-\rho^2)}}$$

Where: η_k are i.i.d. of mean 0 and variance σ_η. In most encountered situations, we can apply the following law of large number to $\bar{w_{i,t}}$.

Law of large number for independent non-identically distributed random variables: Let (X_i) be a sequence of independent random variables such that $\mathbb{E}[|X_i|^{1+\delta}] < \infty$ for some $\delta > 0$ and all i. Then, almost surely,

$$\frac{1}{N}\sum_{i=1}^{N} X_i - \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}(X_i) \xrightarrow[N\to\infty]{} 0$$

This is enough to conclude as $\mathrm{E}(\bar{w}_{i,t}) = \mathrm{E}(w_{i,t}) = \sigma_\varepsilon^2$