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

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

October 9, 2019

Version 1.0.1 - comments are welcome !

Abstract

This note discusses the estimation of the parameters of fixed-effects regression with a AR(1) perturbations. It notably highlights that current estimation procedures fails to take into account possible missingness of the data, with three consequences. Firstly, the variance of the perturbations is currently badly estimated. This note proposes a method to correctly estimate it. Secondly, though it demonstrates that the estimation of the β coefficient is consistent in the case of an unbalanced panel, this demonstration highlights that the proper convergence relies on assumptions akin to a missing-at-random hypothesis, which Monte Carlo simulations confirm. Thirdly and least importantly, it highlights that a correction is necessary in the unbalanced case to make the constant met the usual convention that it makes the mean of the fixed-effects null.

[*]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.* (2019b). 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^2_ε.

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 al.* (2019a), we proposed new estimators for the auto-correlation coefficient. In the particular case of random-effects model (that is: ν_i is exogenous to the covariables of the models) and a positive auto-correlation parameter,[2] it enables to have consistent or asymptotically consistent estimators for all parameters of the model when there are enough individuals with at least two consecutive observations and at least three observations.

We now turn to the more general case when no hypothesis is made regarding the exogeneity of the fixed effects. First we remind the properties of the current estimation method (for instance adopted by the Stata software **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. Second, a consistent estimator of the variance of the perturbations is proposed. Third, we illustrate the above assertions though MonteCarlo simulations that compare the respective performances of the (more specific) random-effects estimators and of the (more general) fixed-effects estimators. We eventually conclude.

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

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

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 al.*, 2019a). 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-effect 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 show that this if fine for the estimation of the variance of the fixed-effects ; that (save for the constant in unbalanced panel, and conditional to making hypotheses akin to a missing-at-random hypothesis) this is also fine for the estimation of the β coefficient ; but that, save for the balanced panel, this does not correctly estimate the variance of the perturbations in the general case.

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:

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

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 as each is a sum of uncorrelated $\eta_{i,t}$.

The current estimation method (see Stata - xtxtregar / 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^{**} is then performed.

Consistent estimation of the β:
It is trivial to see that this 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.

Most interestingly, we show that **it also provides a consistent estimator of** β **in the unbalanced case**, which is not trivial (see Annex A for a demonstration), **under conditions observed under a missing-at-random hypothesis**. It implies that, even in the case of unbalanced panels, β is well estimated under the hypothesis that the fact to be missing is not correlated with the covariates of the model. Note that we provide in Section 4 Monte-Carlo simulations where β is badly

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

estimated when this hypothesis is not respected.

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

An imprecise estimation of the fixed effects ν_i: As such, this method enables also to estimate (although there are not centered around zero):

$$\hat{\nu_i} = y^*_{i,t} - (x'_{i,t}\beta)^* \tag{7}$$

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.

2.2 The estimation of the variance of the perturbations

We now focus on the current estimation of the variance of the perturbations σ_η. The current estimation method takes as an estimator of σ_η the empirical variance of $u^*_{i,t}$. In the unbalanced case, there is no reason why it would yield a consistent estimator of σ_η. Monte Carlo simulation in Section 4 clearly shows that the corresponding estimates can be far away from the true value of σ_η.

It however avers that in the balanced case, it yields a consistent estimator of σ_η. We show below why it is the case

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

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

[4]This convention is usual in Stata.

We remind equation (2).

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

This enables to conclude as we easily get that:

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

3 A new estimator of the variance of the perturbations

A naive estimator of σ_ϵ can be obtained by explicitely computing $\epsilon_{i,t} := y_{i,t} - x'_{i,t}\beta - \nu_i$; considering its empirical variance, and multiplying it by $(1-\rho^2)$ to get the empirical variance of the σ_ν, which can be used as an estimator of σ_ν. It however yields an imprecise estimator of σ_ν, as it relies on the estimation of the c_i. We therefore propose an estimator that is not based on the estimation of the c_i.

To do so, we define:

$$\tilde{y_{i,t}} = y_{i,t} - x'_{i,t}\beta \tag{8}$$

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:

$$\tilde{y_{i,t}} = y_{i,t} - x'_{i,t}\beta = \nu_i + u_{i,t} \tag{9}$$

We define :

$$\tilde{\tilde{y}}_{i,t(i,j)} = \tilde{y}_{i,t(i,j)} - \tilde{y}_{i,t(i,j-1)} = u_{i,t(i,j)} - u_{i,t(i,j-1)} \tag{10}$$

By successively applying equation (2), it comes:

$$\tilde{\tilde{y}}_{i,t(i,j)} = (\rho^{t(i,j)-t(i,j-1)} - 1)u_{i,t(i,j-1)} + \sum_{k=0}^{t(i,j)-t(i,j-1)-1} \rho^k \eta_{i,t(i,j)-k} \tag{11}$$

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

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

Thus:

$$\sigma_\eta^2 = \frac{E(\tilde{\tilde{y}}_{i,t(i,j)}{}^2)}{\frac{(1-\rho^{t(i,j)-t(i,j-1)})^2}{(1-\rho^2)} + \frac{(1-\rho^{2(t(i,j)-t(i,j-1))})}{(1-\rho^2)}} \tag{12}$$

We have obtained:

$$\sigma_\eta^2 = E(w_{i,t}) \tag{13}$$

where:

$$w_{i,t} = \frac{\tilde{\tilde{y}}_{i,t(i,j)}{}^2}{\frac{(1-\rho^{t(i,j)-t(i,j-1)})^2}{(1-\rho^2)} + \frac{(1-\rho^{2(t(i,j)-t(i,j-1))})}{(1-\rho^2)}} \tag{14}$$

By taking the grand average over all values of the term within the expectancy, one gets a consistent estimate of σ_η^2.

$$\hat{\sigma_\eta^2} = \frac{1}{N}\sum_1^N w_{i,t} \tag{15}$$

4 Monte Carlo simulations

To illustrate the above theoretical part, we consider 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 are already used by Cazenave-Lacroutz *et al.* (2019a), enabling comparability. [5]

4.1 In the random-effect design

First, we consider cases where the fixed effects ν_i are exogeneous from the covariables x.[6] This is a very particular case, where a random-effects model can also be applied. This allows us to compare the estimation advantages of making the assumption of a random-effects model (when suitable) rather than the more general fixed-effects model.

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

[6]Typically, we consider as covariables 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 model and the random-effects model are able to provide a correct estimation of the variance of the perturbations (see Table 1). However, due to the limited number of periods observed per individual, the estimation of the variance of the fixed effects σ_ν is biased in the fixed-effects model, but unbiased in the random-effect models.[7]

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

	Fixed-effects model			Random-effects model		
	ρ	σ_ε	σ_ν	ρ	σ_ε	σ_ν
true values	0.6	0.3	0. 35	0.6	0.3	0. 35
with true ρ and xtregar	.6 (0)	.621*** (4.9e-03)	.447*** (.013)	.6 (0)	.3 (4.2e-03)	.356 (.016)
with true ρ and corrected xtregar	.6 (0)	.301 (3.5e-03)	.447*** (.013)	.6 (0)	.3 (4.2e-03)	.356 (.016)
with ρ_{BFN} and corrected xtregar	.581 (.04)	.299 (3.0e-03)	.447*** (.013)	.581 (.04)	.3 (3.9e-03)	.359 (.023)

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: Corrected xtregar applies only to the estimation of σ_ϵ in the Fixed-effects model.

In Table 2, we consider the same simulations, but we increase the number of period to $T = 100$ (rather than $T = 10$). Accordingly with our above interpretations, this yields estimates of the variance of the fixed-effects σ_ν that are no longer significantly different from its true value in the general case of the fixed-effects model.

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

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

	Fixed-effects model			Random-effects model		
	ρ	σ_ε	σ_ν	ρ	σ_ε	σ_ν
true values	0.6	0.3	0. 35	0.6	0.3	0. 35
with true ρ and xtregar	.6 (0)	.678*** (2.8e-03)	.36 (.01)	.6 (0)	.3 (1.1e-03)	.35 (9.9e-03)
with true ρ and corrected xtregar	.6 (0)	.3 (1.2e-03)	.36 (.01)	.6 (0)	.3 (1.1e-03)	.35 (9.9e-03)
with ρ_{BFN} and corrected xtregar	.601 (4.6e-03)	.3 (1.1e-03)	.36 (.01)	.601 (4.6e-03)	.3 (1.2e-03)	.35 (9.9e-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*.

Note 2: Corrected xtregar applies only to the estimation of σ_ϵ in the Fixed-effects model.

4.2 In the general case

We implement two changes in regard with the simulations presented in Table 1.
First, we do not consider any longer the specific case when the fixed effects are exogeneous to the covariates.[8] Hence, we no longer present the random-effects model, as it is based on this hypothesis.
Second, while the data were missing at random, we deviate from this missing pattern. In the "Non missing at random", missingness is based on the value taken by the covariable of the model. As shown in Table 3, this yields a coefficient for the covariable that is significantly different from its true value, which is not observed in the Missing-at-random case.

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 (0)	.621*** (4.9e-03)	.442*** (.019)	3 (.012)	.6 (0)	.458*** (.017)	.377 (.021)	2.95*** (.013)
with true ρ and corrected xtregar	.6 (0)	.301 (3.4e-03)	.442*** (.019)	3 (.012)	.6 (0)	.308 (4.9e-03)	.377 (.021)	2.95*** (.013)
with ρ_{BFN} and corrected xtregar	.581 (.04)	.299 (2.9e-03)	.442*** (.019)	3 (.012)	.581 (.04)	.307 (5.1e-03)	.379 (.023)	2.95*** (.016)

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: *Corrected xtregar* applies only to the estimation of σ_ϵ. *covar* is an independent variable that is the sum of a random term and the fixed-effect.

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

5 Conclusion

Whereas Baltagi et Wu (1999) proposed a way of consistently estimating the parameters of a random-effects regression with AR(1) perturbations[9], 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, followed by a demeaning procedure. We show that this procedure is perfectly adapted in the very particular case of balanced panels, but not necessarily in the more common case of unbalanced panel.

Fortunately for the various papers that used such procedure in the past, under the quite general *missing-at-random* hypothesis, it yields a consistent estimation of the β parameter in the unbalanced case as well. The constant does not *a priori* respects the convention that it makes the mean of the fixed effects null, but the constant is usually not an object of interest *per se*. More importantly for some applications (e.g. for simulations following the estimation), the variance of the perturbation was not correctly 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. Under general hypotheses, all parameters of this type of model are thus consistently estimated.

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. (2019a). The estimation of the autocorrelation coefficient in panel data models with ar(1) disturbances.

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

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

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

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.

Annexes

A Consistency of $\hat{\beta}$

We prove here 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 some equalities hold. According to Lemma 2, under the missing at random hypothesis, these equalities are asymptotically respected for N large enough. This enables to conclude.

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

$$\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}} = (1 - \frac{n_i - 1}{\sum_{i=1}^{N}(n_i - 1)}) \frac{1}{n_i - 1} \sum_{l=2}^{n_i} \frac{1-\rho^{t_{il}-t_{il-1}}}{\sqrt{1-\rho^{2(t_{il}-t_{il-1})}}} \tag{16}$$

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

.

First, we compute

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

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

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

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

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

Then, condition $\mathbb{E}(x^{**}_{it_{ij}}\nu^{**}_{it_{ij}}) = 0$ holds for all $2 \leq j \leq n_i$ if

$$\frac{1-\rho^{t_{ij}-t_{ij-1}}}{\sqrt{1-\rho^{2(t_{ij}-t_{ij-1})}}} = (1-\frac{n_i-1}{\sum_{i=1}^{N}(n_i-1)})\frac{1}{n_i-1}\sum_{l=2}^{n_i}\frac{1-\rho^{t_{il}-t_{il-1}}}{\sqrt{1-\rho^{2(t_{il}-t_{il-1})}}} \tag{17}$$

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

This is equation 16

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

(16) holds in expectancy for N large enough.

If condition 16 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 this case, we would have

$$1 = 1 - \frac{n_i - 1}{\sum_{i=1}^{N}(n_i - 1)}$$

which is valid only for N large enough since $n_i \geq 2$ and is bounded. In other words, provided the pattern of the data is fixed, that is, t_{ij}, $2 \leq j \leq n_i$ are given, condition 16 would hold if the increments $t_{ij} - t_{ij-1}$ were all equal. Although such property might seem restrictive, it is satisfied in average. That is, if the data pattern were random, then condition 16 would be valid in expectancy. To see this, we compute

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

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

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

Then, condition 16 holds in expectancy if and only if

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

which is the case when N grows to infinity.