ARTICLE IN PRESS
Journal of Econometrics ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Unequal spacing in dynamic panel data: Identification and estimation[☆]

Yuya Sasaki [*], Yi Xin

Department of Economics, Johns Hopkins University, Wyman Park Building 544E, 3100 Wyman Park Drive, Baltimore, MD 21211, United States

ARTICLE INFO

Article history:
Received 20 May 2015
Received in revised form
23 January 2016
Accepted 17 October 2016
Available online xxxx

JEL classification:
C23

Keywords:
Dynamic panel data
Earnings dynamics
Unequal spacing

ABSTRACT

We propose conditions under which parameters of fixed-effect dynamic models are identified with unequally spaced panel data. Under predeterminedness, weak stationarity, and empirically testable rank conditions, AR(1) parameters are identified given the availability of "two pairs of two consecutive time gaps", which generalizes "two pairs of two consecutive time periods". This result extends to models with multiple covariates, higher-order autoregressions, and partial linearity. Applying our method to the NLS Original Cohorts: Older Men, where personal interviews took place in 1966, 67, and 69, we analyze the earnings dynamics in the old time, and compare the results with more recent ones.

1. Introduction

In economics, numerous empirical questions have been answered through the dynamic panel data model of the form

$$y_{it} = \gamma y_{i,t-1} + \beta x_{it} + \alpha_i + \varepsilon_{it}, \tag{1.1}$$

where y_{it} is an observed state variable, x_{it} is an observed covariate, α_i is an unobserved individual fixed effect, and ε_{it} is an idiosyncratic error. Among others, method-of-moment approaches (e.g., Anderson and Hsiao, 1981; Arellano and Bond, 1991) enjoy practical and theoretical advantages to attract a large group of users. These methods exploit the instrumental orthogonality of the first difference $\varepsilon_{it} - \varepsilon_{i,t-1} = (y_{it} - y_{i,t-1}) - \gamma(y_{i,t-1} - y_{i,t-2}) - \beta(x_{it} - x_{i,t-1})$ as well as other supplementary moment restrictions. As such, they require observation of y_{it} for at least three consecutive time periods (or alternatively two pairs of two consecutive time periods).

Many panel surveys are conducted with unequal time spacing, and may not provide the required set of time periods. For the NLS Original Cohorts: Older Men, for example, personal interviews were conducted in 1966, 67, 69, 71, 76, 81, and 90.[1] This data set contains neither three consecutive time periods nor two pairs of two consecutive time periods. We thus fail to difference out the fixed effect from Eq. (1.1), and cannot directly adapt the aforementioned approaches to construct moment restrictions.

Given that the standard method-of-moment approaches are not generally effective once panel data exhibit unequal time spacing, can we develop similarly useful alternative estimation methods? Through this paper, we answer this question by providing conditions under which parameters (γ, β) of the model (1.1) are identified even if panel data are unequally spaced. In addition to the relatively standard assumptions such as predeterminedness, weak stationarity, and empirically testable rank conditions, we require certain patterns of unequal time spacing for the parameters to be identified. It is also shown that many of the unequally spaced panel data sets from the US and the UK satisfy our requirement of spacing patterns.

We are not the first to study unequally spaced panel data. Rosner and Munoz (1988) use linear interpolation to approximate missing data for dynamic panel models. Jones and Boadi-Boateng (1991) take the parametric maximum likelihood solution for static panel models with serial correlation. Baltagi and Wu (1999) propose a feasible GLS procedure for static panel models with

☆ An associate editor and three anonymous referees contributed much to this paper. We are indebted to them. We also benefited from very useful comments obtained in New York Camp Econometrics 2014, AMES 2014, and International Panel Data Conference 2014. The usual disclaimer applies.
* Corresponding author.
E-mail addresses: sasaki@jhu.edu (Y. Sasaki), yxin4@jhu.edu (Y. Xin).

1 They conducted mail or telephone interviews in 1968, 73, 75, 78, 80, and 83, but responses through different media of communication should be carefully distinguished for survey analysis.

http://dx.doi.org/10.1016/j.jeconom.2016.10.002
0304-4076/

Please cite this article in press as: Sasaki, Y., Xin, Y., Unequal spacing in dynamic panel data: Identification and estimation. Journal of Econometrics (2016), http://dx.doi.org/10.1016/j.jeconom.2016.10.002

serial correlation. McKenzie (2001) shows consistent estimation of dynamic (pseudo) panel models, but this method requires observation of covariates in the missing time periods. Millimet and McDonough (2013) apply a variety of estimation methods and report their finite sample performances. They also provide a comprehensive survey of panel data with unequal spacing in Australia, Canada, France, Japan, UK and USA. A more recent paper by Pacini and Windmeijer (2015) considers the AR(1) model with randomly missing outcomes, where they assume that covariates are always observed or are absent from the model.

This paper differs from these preceding papers in terms of the following six points. First, most importantly, we show identification, and specifically propose general spacing patterns as sufficient conditions for identification. Second, we deal with dynamic models which exhibit more complications than static models. Third, parametric distributional assumptions are not imposed. Fourth, our approach does not rely on interpolation or imputation. Fifth, our method can allow for arbitrarily correlated covariates and does not require partial observation in missing time periods. Sixth, our model can allow for arbitrary correlation among the observed state, the unobserved fixed effect, and the observed covariates.

With all these advantages, we admit that our identification result is based on a non-trivial set of assumptions. As mentioned earlier, we assume predeterminedness and weak stationarity. While predeterminedness is often innocuous in applications, the weak stationarity can be restrictive in some applications. We discuss advantages and disadvantages of this assumption. Our rank condition is empirically testable, and can also be handled by the existing methods of weak-rank-robust inference.

Our key requirement for identification is the availability of "two pairs of two consecutive time gaps", which is a generalization of "two pairs of two consecutive time periods". None of the preceding papers proposes such general spacing patterns as sufficient (or necessary) condition for identification. This requirement is satisfied with time gaps $\{0, 1\}$ and $\{1, 2\}$ serving as two pairs of two consecutive time gaps for the NLS Original Cohorts: Older Men, which we picked as an example earlier. This paper contributes to the body of our knowledge and provides a guidance to practitioners by formally ensuring identification of dynamic fixed-effect models under the stylized patterns of unequally spaced panel data.

2. A basic model

We first fix index notations for unequally spaced panel data. Let T be the set of all observed time periods. Define the set of survey gaps by $\mathcal{T} = \{|t_1 - t_2| : t_1, t_2 \in T\}$. Also define the set of gap-associated survey years by $T(\tau) = \{t \in T : t + \tau \in T\}$ for each gap $\tau \in \mathcal{T}$, and let $T(\tau) = \emptyset$ if $\tau \notin \mathcal{T}$. For the NLS Original Cohorts: Older Men, introduced in the previous section, personal interviews were conducted in 1966, 67, 69, 71, 76, 81, and 90. In this case, we have $T = \{66, 67, 69, 71, 76, 81, 90\}$, $\mathcal{T} = \{0, 1, 2, 3, 4, 5, 7, 9, 10, 12, 14, 15, 19, 21, 23, 24\}$, $T(0) = T$, $T(1) = \{66\}$, $T(2) = \{67, 69\}$, $T(3) = \{66\}$, $T(4) = \{67\}$, $T(5) = \{66, 71, 76\}$, and so on.

Let us first consider the following simple first-order autoregressive model for illustration

$$y_{it} = \gamma y_{i,t-1} + \beta x_{it} + \alpha_i + \varepsilon_{it} \tag{2.1}$$

where y_{it} is an observed state variable, x_{it} is an observed covariate, α_i is an unobserved individual fixed effect, and ε_{it} is an unobserved idiosyncratic shock for individual i at period t. This baseline model has two parameters, γ and β. Our model, allowing for an arbitrary correlation among the observed state, the unobserved fixed effect, and the observed covariates, generalizes the model of Millimet and McDonough (2013), where x_{it} is required to be strictly exogenous and serially uncorrelated for consistency, and the model of Pacini and Windmeijer (2015), where covariates treated separately for consistency. With this said, we will invoke a time invariance assumption below that was not used by them. We introduce the short-hand notation, $\mathrm{E}_i(\cdot) := \mathrm{E}(\cdot|\alpha_i)$, for the expectation conditional on individual i's specific heterogeneity. The dynamic process (2.1) is equipped with the following set of model assumptions.

Assumption 1 (*Predeterminedness*). $E_i(y_{it}\varepsilon_{is}) = 0$ and $E_i(x_{it}\varepsilon_{is}) = 0$ whenever $s > t$.

Assumption 2 (*Time Invariance*). For each individual $i = 1, 2, \ldots, N$ and $\tau \geq 0$:

(i) $\mathrm{E}_i(y_{i1}y_{i,1+\tau}) = \cdots = \mathrm{E}_i(y_{it}y_{i,t+\tau}) = \cdots = \mathrm{E}_i(y_{i,T-\tau}y_{iT})$.
(ii) $\mathrm{E}_i(x_{i1}x_{i,1+\tau}) = \cdots = \mathrm{E}_i(x_{it}x_{i,t+\tau}) = \cdots = \mathrm{E}_i(x_{i,T-\tau}x_{iT})$.
(iii) $\mathrm{E}_i(y_{i1}x_{i,1+\tau}) = \cdots = \mathrm{E}_i(y_{it}x_{i,t+\tau}) = \cdots = \mathrm{E}_i(y_{i,T-\tau}x_{iT})$.
(iv) $\mathrm{E}_i(x_{i1}y_{i,1+\tau}) = \cdots = \mathrm{E}_i(x_{it}y_{i,t+\tau}) = \cdots = \mathrm{E}_i(x_{i,T-\tau}y_{iT})$.

The predeterminedness assumption based on E_i requires the individual-level moment equalities $\mathrm{E}(y_{it}\varepsilon_{is}|\alpha_i) = \mathrm{E}(x_{it}\varepsilon_{is}|\alpha_i) = 0$ to hold for any $s > t$. By the law of iterated expectation, this implies $\mathrm{E}(y_{it}\varepsilon_{is}) = \mathrm{E}(x_{it}\varepsilon_{is}) = 0$ for $s > t$. Unlike the usual exogeneity conditions assumed in the panel data literature, our assumption is based only on the individual-level moments $\mathrm{E}_i(\cdot)$ and the moment equalities need not hold for $s = t$, but it suffices for the purpose of identification as formally argued below. If we substitute stronger conditions such as $\mathrm{E}_i(x_{it}\varepsilon_{is}) = 0$ for all $s \geq t$ or the strict exogeneity, then with more data availability we can possibly gain identification power and efficiency through more moment conditions. We also remark that the empirical testability of the predeterminedness in general requires exogenous instruments.

Conditions in Assumption 2 are satisfied when the variables are weakly stationary. In light of this assumption, we can define auxiliary random variables $Z_{i\tau} := \mathrm{E}_i(y_{it}y_{i,t+\tau})$, $z_{i\tau} := \mathrm{E}_i(x_{it}x_{i,t+\tau})$, $\zeta_{i\tau} := \mathrm{E}_i(y_{it}x_{i,t+\tau})$ and $\zeta_{i,-\tau} := \mathrm{E}_i(x_{it}y_{i,t+\tau})$, which do not depend on t.[2] For a primitive structural model that sufficiently satisfies this weak stationarity assumption, one can consider the linear VAR structure for $(y_{it}, x_{i,t+1})$ with sub-unit coefficient restrictions for example, where one of the two equations is (2.1). Note that our assumption does not require the initial observations to be generated from the stationary distribution.[3] Compared to the existing literature (e.g., Millimet and McDonough, 2013), the time invariance assumption is the key to our result. This assumption has both advantages and disadvantages. On one hand, it facilitates the identification that we establish under unequal spacing and arbitrary correlation among the observed state, the unobserved fixed effect, and the observed covariates that are neither strictly exogenous nor serially uncorrelated. It also allows us to gain identification power and efficiency in traditional settings—see Example 4 for detailed discussions. On the other hand, this assumption can be too restrictive for certain applications, particularly with state variables that grow or accumulate over time. In addition, time-series heteroskedasticity, which is usually allowed in existing estimation methods for equally spaced panels, is excluded by this assumption.

2 We introduce these auxiliary variables only for the sake of making equations shorter, but they are not necessary for the substance of our formal discussions.

3 To see this, observe that Assumption 2 allows the initial observations in a pure AR(1) model to take the general form of $y_{i0} = \delta\left(\frac{\alpha_i}{1-\gamma}\right) + \varepsilon_{i0}$ with $\varepsilon_{i0} \sim N(0, \frac{\sigma_\varepsilon^2}{1-\gamma^2})$. Within this general framework, on the other hand, a stationary initial distribution specifically requires $\delta = 1$.

Please cite this article in press as: Sasaki, Y., Xin, Y., Unequal spacing in dynamic panel data: Identification and estimation. Journal of Econometrics (2016), http://dx.doi.org/10.1016/j.jeconom.2016.10.002