ARTICLE IN PRESS
Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Median-based estimation of dynamic panel models with fixed effects[☆]

Geert Dhaene[a,*], Yu Zhu[b]

[a] KU Leuven, Department of Economics, Naamsestraat 69, B-3000 Leuven, Belgium
[b] Bank of Canada, 234 Laurier Avenue West, Ottawa, Ontario, Canada

ARTICLE INFO

Article history:
Received 11 November 2015
Received in revised form 11 April 2016
Accepted 2 May 2016
Available online xxxx

Keywords:
Additive outliers
Breakdown point
Dynamic panel data
Fixed effects
Influence function
Robustness

ABSTRACT

Outlier-robust estimators are proposed for linear dynamic fixed-effect panel data models where the number of observations is large and the number of time periods is small. In the simple setting of estimating the AR(1) coefficient from stationary Gaussian panel data, the estimator is (a linear transformation of) the median ratio of adjacent first-differenced data pairs. Its influence function is bounded under contamination by independent or patched additive outliers. The influence function and the gross-error sensitivity are derived. When there are independent additive outliers, the estimator is asymptotically biased towards 0, but its sign remains correct and it has a reasonably high breakdown point. When there are patched additive outliers with point mass distribution, the asymptotic bias is upward in nearly all cases; breakdown towards 1 can occur; and the associated breakdown point increases with the patch length.

1. Introduction

Suppose we have N units of observation for which there are $T \geq 3$ repeated measurements of a scalar variable y. That is, we observe y_{it} for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. Assume that the y_{it} are generated by heterogeneous Gaussian AR(1) processes with common autoregressive parameter ρ, i.e.,

$$y_{it} = \alpha_i + \rho y_{it-1} + \varepsilon_{it}, \quad \varepsilon_{it} \sim N\left(0, \sigma_i^2\right), \tag{1.1}$$

for $i = 1, \ldots, N$ and $t = 2, \ldots, T$, with errors ε_{it} assumed to be independent across i and t, and $-1 < \rho \leq 1$. The fixed effects α_i and the error variances $\sigma_i^2 > 0$ are nuisance parameters. The pairs (α_i, σ_i^2) may be fixed or stochastic but are assumed to be predetermined relative to y_{it} $(t = 1, \ldots, T)$. The initial values y_{i1} are assumed to be independent across i and drawn from the stationary distributions when $|\rho| < 1$, i.e.,

$$y_{i1} \sim N\left(\frac{\alpha_i}{1-\rho}, \frac{\sigma_i^2}{1-\rho^2}\right) \tag{1.2}$$

for $i = 1, \ldots, N$, and may be any values (fixed or stochastic) when $\rho = 1$. We address the problem of estimating ρ in such a way that the estimator is robust against data contamination and also consistent as $N \to \infty$ with fixed T when there is no contamination. Some generalizations of the basic model (1.1)–(1.2) are also considered.

☆ The views expressed here are those of the authors and do not necessarily reflect the position of the Bank of Canada.

* Corresponding author. Fax: +32 16326796.
E-mail addresses: geert.dhaene@kuleuven.be (G. Dhaene), zhuy@bank-banque-canada.ca (Y. Zhu).

http://dx.doi.org/10.1016/j.csda.2016.05.021

Please cite this article in press as: Dhaene, G., Zhu, Y., Median-based estimation of dynamic panel models with fixed effects. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.05.021

The model $y_{it} = \alpha_i + \rho y_{it-1} + \varepsilon_{it}$ is the simplest of a range of dynamic fixed-effect panel models including models with covariates and higher order dynamics. Such data and models have a long history of applied research starting with the work of Balestra and Nerlove (1966). In applications, N is often large and T small (say, $N \geq 1000$ and $T \leq 10$), so it is natural then to consider asymptotics where $N \to \infty$ while T remains fixed. As is well known, the least-squares estimator of ρ is not consistent as $N \to \infty$ with T fixed. Least-squares, here, is the within-group estimator $\sum_i \sum_{t\geq 2}(y_{it} - \bar{y}_i)(y_{it-1} - \bar{y}_{i-})/\sum_i \sum_{t\geq 2}(y_{it-1} - \bar{y}_{i-})^2$, where $\bar{y}_i$ and $\bar{y}_{i-}$ are the means of y_{it} and y_{it-1} over $t \geq 2$. Nickell (1981) derived its large N, fixed T bias, which is due to an incidental-parameter problem first described, in a more general setting, by Neyman and Scott (1948). When T is small, the bias is particularly large, and the general reaction has been to move away from least-squares to GMM estimators such as those proposed by Anderson and Hsiao (1981, 1982), Holtz-Eakin et al. (1988), Arellano and Bond (1991), Ahn and Schmidt (1995), and Blundell and Bond (1998). These estimators are large N, fixed T consistent but were not designed to be outlier-robust. They are, in fact, highly sensitive to even a small number of outliers. Lucas et al. (2007) show that the Arellano–Bond estimator has an unbounded influence function, and the same can be shown for the other GMM estimators. Recently, there has been a return to the estimation of dynamic fixed-effect panel models using likelihood and least-squares methods, but in a manner that resolves the incidental-parameter problem; see, e.g., Hsiao et al. (2002), Lancaster (2002), Moreira (2009), and Han et al. (2014). However, as with GMM, these estimators are large N, fixed T consistent but not robust to outliers.

We propose an estimator of ρ that is consistent (as $N \to \infty$, for any $T \geq 3$) and has attractive robustness properties against data contamination. The basic version of the estimator is $\widehat{\rho} = 1 + 2\widehat{r}$, where $\widehat{r}$ is the median of the ratios $(y_{it} - y_{it-1})/(y_{it-1} - y_{it-2})$, and a slight variation of $\widehat{\rho}$ is exactly unbiased for any $N \geq 3$ and $T \geq 3$. Intuitively, $\widehat{\rho}$ is a highly robust estimator because it is a median. It has a bounded influence function and a reasonably high breakdown point when the data are contaminated by independent or patched additive outliers. Under independent additive outlier contamination, $\widehat{\rho}$ is asymptotically biased towards zero but is sign-robust (i.e., its sign remains correct) regardless of the contamination rate and the outlier distribution. The breakdown point towards 0, for which we derive an upper and a lower bound, depends on ρ. When there are patched additive outliers with point-mass distribution, the bias of $\widehat{\rho}$ is upward except in certain cases where ρ and the contamination rate are large. We give an account of these and other robustness properties in Section 3 by applying concepts from robust time series statistics (in particular, Martin and Yohai, 1986) to the fixed-effect panel data setting. In line with most of the robustness literature, our approach relies on the normality assumption, but it can, at least in principle, be extended to the weaker assumption that the joint distribution of $y_{it} - y_{it-1},\ t = 2, \ldots, T$, is elliptically contoured. We address this briefly in Section 4, together with the inclusion of covariates and higher order dynamics in (1.1). In Section 5 we compare $\widehat{\rho}$ with other estimators in simulations with and without data contamination. Section 6 concludes. Proofs are given in an Appendix.

The proposed estimator is related to an estimator of the autocorrelation of a Gaussian zero-mean AR(1) process suggested by Hurwicz (1950). It can be viewed as an extension of Hurwicz' estimator to the fixed-effect panel data setting accommodating non-zero means. Being a median, it also bears similarity to certain regression slope estimators under random sampling, in particular the median-type GM estimator of Brown and Mood (1951), the median of pairwise slopes of Theil (1950) and Sen (1968), and the repeated median of pairwise slopes of Siegel (1982). Adrover and Zamar (2004) called these regression estimators "median based" and investigated their maximum asymptotic bias, concluding that their bias-robustness properties are very satisfactory.

At a general level, we emphasize the need for further development of outlier-robust methods for panel data analysis. It is generally acknowledged that standard inferential methods tend to be fragile to outliers, and commonly used panel data methods are no exception. Furthermore, the abundance of information in large panel data sets naturally suggests moving towards methods that are less fragile to outliers, even though this typically entails a loss of statistical efficiency. At present, the body of literature on outlier-robust panel data methods is still small, presumably because it is challenging to deal with both unobserved heterogeneity (typically through fixed or random effects) and outliers simultaneously. The earliest work on robustness in panel data was by Wagenvoort and Waldmann (2002), who studied bounded-influence estimation in a linear static panel data model with endogeneity but without unobserved heterogeneity. Subsequent work by Bramati and Croux (2007), Lucas et al. (2007), Baltagi and Bresson (2012), Aquaro and Čížek (2013, 2014), and Čížek and Aquaro (2015) developed bounded-influence and high-breakdown estimators in linear static and dynamic panel data models with fixed effects. Aquaro and Čížek (2014) and Čížek and Aquaro (2015) build on results presented in this paper.

2. Median-based estimators

Hurwicz (1950) observed that, in the time series model $y_t = \rho y_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim N\left(0, \sigma^2\right)$, every ratio y_t/y_{t-1} is median-unbiased for ρ and conjectured that the median of those ratios is median-unbiased for ρ. Zieliński (1999) proved Hurwicz's conjecture, provided that the median of, say, $x_1, \ldots, x_n$ with order statistics $x_{(1)} \leq \cdots \leq x_{(n)}$, is defined as in Zieliński (1995), viz.

$$\mathrm{med}_Z\{x_1, \ldots, x_n\} = \begin{cases} x_{(k)} & \text{if } n = 2k - 1, \\ Dx_{(k)} + (1 - D)x_{(k+1)} & \text{if } n = 2k, \end{cases}$$