Journal of Econometrics 194 (2016) 57–75

Contents lists available at ScienceDirect

Journal of Econometrics

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

Identification of panel data models with endogenous censoring[☆]

Shakeeb Khan[a,*], Maria Ponomareva[b], Elie Tamer[c]

[a] *Department of Economics, Duke University, Durham, NC, USA*
[b] *Department of Economics, Northern Illinois University, DeKalb, IL, USA*
[c] *Department of Economics, Harvard University, Cambridge, MA, USA*

ARTICLE INFO

Article history:
Received 29 September 2011
Received in revised form
3 December 2015
Accepted 4 January 2016
Available online 8 April 2016

Keywords:
Panel data
Dependent censoring
Partial identification
Sharp set
Treatment effect

ABSTRACT

We study inference on parameters in linear panel data models when outcomes are censored. We allow the censoring to depend on both observable and unobservable variables in arbitrary ways. Generally, these models are set identified and the main contribution of this paper is to derive and characterize the identified sets under general conditions. Our main characterization theorems show that every parameter in the sharp set – and only those parameters – can generate the observed data under the maintained assumptions. In particular, we consider two separate sets of assumptions (2 models): the first uses stationarity on the unobserved disturbance terms. The second is a nonstationary model with a conditional independence restriction. Based on the characterizations of the identified sets, we provide an inference procedure that is shown to yield valid confidence sets based on inverting stochastic dominance tests. We also show how our results extend to empirically interesting dynamic versions of the model with both lagged observed outcomes, lagged indicators, and models with factor loads. In addition, we provide sufficient conditions for point identification in terms of support conditions. The paper then examines the size of the identified sets in particular designs, and a Monte Carlo exercise shows reasonable small sample performance of our procedures. We also apply our inference approach to two empirical illustrations that link endogenous censoring to treatment effects models.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

We consider the problem of inference on the k-dimensional parameter $\beta \in B \subset \mathbf{R}^k$ in the panel data model

$$y_{it}^* = \alpha_i + x_{it}'\beta + \epsilon_{it}, \quad t = 1, \ldots, T \quad i = 1, \ldots, N$$

where α_i is an individual specific and time-independent fixed (or random) effect that is allowed to be correlated with *both* $\mathbf{x}_i = (x_{i1}, \ldots, x_{iT})$ and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iT})$. The outcome variable, y_{it}^*, is only observed when it is greater than a censoring variable c_{it}. The censoring variable c_{it} itself is observed only when it exceeds y_{it}^*. The censoring variable $\mathbf{c}_i = (c_{i1}, \ldots, c_{iT})$ is allowed to depend on $\boldsymbol{\epsilon}_i$ in an arbitrary way. We summarize this as follows:

$$\begin{array}{ll} \text{we observe for } i: & \left(y_{it} = \max(y_{it}^*, c_{it}),\ 1[y_{it}^* \geq c_{it}],\ x_{it}\right) \\ \quad t = 1, \ldots, T & \\ \qquad\quad \text{where} & \boldsymbol{\epsilon}_i \not\perp \mathbf{c}_i | \mathbf{x}_i. \end{array}$$

The presence of this *endogenous* censoring represents a challenge for existing methods that are used for correcting for censoring since these methods usually assume that $\mathbf{c}_i$ is either observed or (conditionally) independent of the errors. There, the observed censoring is motivated via design or data limitation issue (such as top-coding), and hence is assumed independent of the outcome. Here, the starting point is we want this censored variable c_{it} to be on equal footing as the outcome and so allow it to be arbitrarily correlated with y_{it}^* (but also accommodate fixed and independent censoring[1]). Allowing for endogenous censoring is critical in data sets where the censoring can be a function of unobservables that

[☆] We thank Sofya Budoneva, Marcelo Ochoa and Yichong Zhang for research assistance on this project, seminar participants at various institutions for comments, and the National Science Foundation for research support. We also thank the editor and the reviewers for constructive comments on the paper.

[*] Corresponding author.
E-mail addresses: shakeebk@duke.edu (S. Khan), mponomareva@niu.edu (M. Ponomareva), elietamer@fas.harvard.edu (E. Tamer).

[1] In the cross sectional setting this model is popular in duration analysis, as it relates to the Accelerated Failure Time (AFT) model. See, e.g Khan and Tamer (2009) for more on this for cross sectional data. In the panel data setting considered in this paper, t does not refer to the time period, but the spell in question.

http://dx.doi.org/10.1016/j.jeconom.2016.01.010
0304-4076/© 2016 Elsevier B.V. All rights reserved.

may be correlated with outcomes. This increases the set of models that are covered to include competing risks models, switching regression like models, and duration models with attrition[2] that are important in applied work.[3]

Generally, point identification conditions in nonlinear panel data[4] models with fixed effects are often strong, partly since simple differencing techniques, used in linear models, are not available when the model is nonlinear in the unobserved individual specific variable. So, typical point identification strategies have relied on distributional assumptions, and/or support conditions that are problem specific that often rule out economically relevant models and behaviors. This has motivated a complementary approach to inference in these models that recognizes the fact that though point identification might not be possible under weaker assumptions, these models do contain nontrivial information about β. So, instead of looking for conditions under which point identification is guaranteed, we posit a model for the data generating process and then analyze the question of what information this model has about β given the observed data.

The challenge in this *bounds approach* to identification analysis is to consider all the information in the data and the model: that is, find the *tightest* set or *sharp set* that contains the observationally equivalent parameter values. Our approach posits a model first, and then asks given this model, what is the sharp set. This is in contrast to the complementary approach based on point identification in which one looks for a model (a set of assumptions) that guarantees point identification under the weakest set of assumptions. The main results in the paper provide characterizations of the sharp sets for β in classes of linear panel models with censoring.

Our censoring mechanism can be seen as a panel extension of the Roy model (or switching regression model) where in every period, one chooses to work in one of two sectors and this decision is based on whether the wage in the one sector is higher than the wage in the other sector. Admittedly, the dynamic decision problem allowed is limited to perfect foresight where future expectations do not play a direct/explicit role. We are also able to show cases that lead to the identification becoming trivial: any possible vector of parameters is consistent with the distribution of observables.[5] Finally, censored models such as ours can be seen as missing or interval outcome models, and were initially considered in the partial identification literature with cross section data.[6]

The recent literature on nonlinear panel data models is extensive and growing. See for example the work of Arellano and Bonhomme (2009), Bester and Hansen (2009), Bonhomme (2012), Chernozhukov et al. (2013), Evdokimov (2010), Graham and Powell (2012) and Hoderlein and White (2012). An important early work is the paper by Honoré (1992) which considers a panel model with fixed censoring. See also the survey in Arellano and Honoré (2001). Most of this work is based on point identification, so it is important to compare the conditions imposed there to ours.

1.1. Comparison to recent point identification results

The closest recent papers with panel data are Hoderlein and White (2012) (HW) and Bonhomme (2012) (SB). HW consider a class of panel data models which allows for both continuous and discrete outcomes. The focus of their results is on point identification of parameters of interest, which can include, but is not limited to regression coefficients and/or average partial effects. Their most general setup, (their) Eq. (2.1), is more general than ours, but their conditions for point identification are stronger. For one, they assume *all* the covariates are continuously distributed. As mentioned in Arellano and Honoré (2001), such a condition generally rules out time dummies. They also require that the unobserved disturbance component be distributed independently from a subset of regressors, conditional on a different set of regressors and the fixed effect (their Assumption A2). This "exclusion" or "partial exogeneity" restriction is reminiscent of, for example, Lewbel (1998), and Honoré and Lewbel (2002), though HW do not impose the large regressor support conditions assumed in those papers. But it is important to note that even with these conditions(continuity and exclusion), the point identification result in HW is "irregular"[7] as they condition on a "thin" set where *all* continuously distributed regressors are identical in consecutive time periods. So, we view HW's results as complementary to ours.

Similarly to HW, SB's paper aims at obtaining point identification at the expense of stronger conditions. In the first part of SB, Section 3, SB considers the finite support case. This includes the assumption of finite support on the individual specific effects. In general (but still finite support) nonlinear models, a condition of non-surjectivity is necessary for the finite dimensional parameter of interest to be point-identified. Non-surjectivity holds in, for example, static binary choice models provided $N_\alpha < 2T$ where N_α is the number of support points of the individual specific effect, and T is the number of time periods. Compared to our paper, we are not restricting the support of the individual effect, and hence are generally only able to attain set identification. SB also contains further interesting results on *local* identification when the support of both observables and unobservables is infinite. SB's results highlight that generally, it is hard to attain (global) point identification in nonlinear panel data models.

1.2. Summary of our models

In this paper the first set of assumptions we employ (**Model 1**) uses *stationarity* on the distribution of $\boldsymbol{\epsilon}_{it}$, but otherwise leaves the error distribution unconstrained (and hence allow for cross sectional heteroskedasticity). Stationarity in nonlinear panel models has been used before in the work of Manski (1987) where it was shown that the binary choice panel model point identifies β under stationarity and a set of support conditions.

The second set of assumptions (**Model 2**) relaxes stationarity but instead imposes independence between $\boldsymbol{\epsilon}_i$ and $\mathbf{x}_i$. This non-stationary setup allows for the distribution of the error terms to vary arbitrarily across time periods. Again, we construct another set of conditional moment inequalities that is shown to

[2] Another interesting example here is the analysis of unemployment duration data in the presence of (right) censoring. Presumably, individuals that drop out of the sample (and hence have censored durations) are different than ones with observed durations and where this difference can be due to unobserved heterogeneity.

[3] A canonical empirical example of this kind of censoring is a wage panel regression with an indicator dummy of whether individual i belongs to a union in time t. The censoring occurs since a union member's nonunion wage in period t is censored but is presumed less than the observed union wage.

[4] For recent developments in the panel data literature in econometrics see Arellano (2003).

[5] A similar example is shown by Rosen (2012) for quantile panel data models with fixed effects and small T. In particular, under a conditional median independence assumption on ϵ_{it}, Rosen (2012) shows that a linear panel model (with no censoring) contains no information on the true parameter β under median restrictions, so that the identified set is the whole parameter space. This happens because no restrictions were made on the *j*oint distribution of ϵ_{i1} and ϵ_{i2}.

[6] Manski and Tamer (2002) considered inference on the slope vector in a linear model with interval outcomes using a partial identification approach. With panel data, Honoré and Tamer (2006) considered bounds on parameters of interest in nonlinear panel models with dynamics.

[7] See Khan and Tamer (2010) for potentially severe consequences of irregular identification in terms of estimation and inference.

Download English Version:

https://daneshyari.com/en/article/5095702

Download Persian Version:

https://daneshyari.com/article/5095702

Daneshyari.com