Journal of Econometrics 184 (2015) 347–360

Contents lists available at ScienceDirect

Journal of Econometrics

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

Inference in semiparametric binary response models with interval data[☆]

Yuanyuan Wan[a,*], Haiqing Xu[b]

[a] Department of Economics, University of Toronto, 150 St. George Street, Max Gluskin House, Toronto, M5S3G7, ON, Canada

[b] Department of Economics, The University of Texas at Austin, United States

ARTICLE INFO

Article history:
Received 12 August 2013
Received in revised form
17 September 2014
Accepted 17 September 2014
Available online 12 October 2014

JEL classification:
C12
C14
C24

Keywords:
Interval data
Semiparametric binary response model
Density weights
$\mathscr{U}$-process

ABSTRACT

This paper studies the semiparametric binary response model with interval data investigated by Manski and Tamer (2002). In this partially identified model, we propose a new estimator based on MT's modified maximum score (MMS) method by introducing density weights to the objective function, which allows us to develop asymptotic properties of the proposed set estimator for inference. We show that the density-weighted MMS estimator converges at a nearly cube-root-n rate. We propose an asymptotically valid inference procedure for the identified region based on subsampling. Monte Carlo experiments provide supports to our inference procedure.

1. Introduction

Interval data is a common feature in empirical research. For example, as an explanatory variable, family income might be measured by a bracket with only upper and lower bounds reported to researchers. Models with interval data have been systematically investigated in a seminal paper by Manski and Tamer (2002, MT). For a semiparametric binary response model with interval data, MT propose a modified maximum score (MMS) set estimator and show its consistency. The convergence rate and other asymptotic properties of the MMS estimator, which are necessary for inference, however are not established. In this paper, we extend MT's method and propose a density-weighted MMS (set-valued) estimator, which allows us to establish the asymptotic properties. Further, we propose an asymptotically valid inference procedure for the identified set. Monte Carlo experiments are used to illustrate the finite sample performance of the proposed estimator and inference procedure.

When one explanatory variable v is not observed but other variables x have been measured precisely, the conditional distribution $\mathbb{P}(y|x, v)$ is unknown in the population. MT suggest to characterize the identification region of model parameters based on $\mathbb{P}(y|x, v_0, v_1)$, where v_0 and v_1 are observed lower/upper bounds of v in the interval data. Instead of modifying the original econometric model and objects of interests, e.g. replacing $\mathbb{P}(y|x, v)$ by $\mathbb{P}(y|x, d)$ where d is a discrete random variable indicating which bracket v belongs to, MT's approach treats the observability of data as a separate issue of modeling and data generating process. Although the observed bounds are less informative than v, they still provide (partial) identification power for the object of interest. MT characterize the sharp identification region for the model parameters and show that their set estimators are consistent.[1] Following that direction, this paper focuses on the interval data issue in

☆ We are grateful to two anonymous referees, an Associate Editor and the Co-Editor for their insightful comments, which helped generate a substantially improved paper. We have greatly benefited from interactions with Jason Abrevaya, Victor Aguirregabiria, Sung Jae Jun, Isabelle Perrigne, Joris Pinkse, and Quang Vuong, and the comments from Tim Armstrong, Jason Blevins, Federico Bugni, Qi Li, George Shoukry, Xun Tang and seminar participants at the 2013 Asia Meeting of Econometric Society, the 2013 China Meeting of Econometric Society and the 9th GNYEC. Wan gratefully acknowledges research support from SSHRC Insight Grant. All errors are ours.

* Corresponding author.
E-mail addresses: yuanyuan.wan@utoronto.ca (Y. Wan), h.xu@austin.utexas.edu (H. Xu).

1 Magnac and Maurin (2008) discuss the identification of the semiparametric binary response model with interval data when additional instrumental variables are available.

http://dx.doi.org/10.1016/j.jeconom.2014.09.009

a semiparametric binary response model and provides an effective inference procedure for the partially identified parameters.

The issue of interval data also arises in estimating game theoretic models, where some equilibrium variables (e.g., equilibrium beliefs) are not observed but we could possibly derive their estimable upper/lower bounds from equilibrium conditions and model restrictions. For example, in a 2-by-2 game of incomplete information with correlated types, Wan and Xu (2012) show that each player's equilibrium strategy can be represented as a binary response model, in which one of the explanatory variables, the equilibrium belief on the rival's choice, is unknown to researchers and bounded by some nonparametrically estimable functions.

In this paper, we extend MT's MMS method by introducing density weights to the objective function for their MMS estimation. The weighting does not change the identification region of parameters of interest, but allows us to obtain a sample objective function in a $\mathscr{U}$-process form. We further extend (Kim and Pollard, 1990)'s results on the asymptotic properties for maximum score point estimator to our setting and establish a set of conditions under which our density-weighted MMS estimator is nearly cube-root-n consistent.

Moreover, we follow Chernozhukov et al. (2007) and construct confidence regions for the partially identified set as level sets of the sample objective function. Abrevaya and Huang (2005) show that the bootstrap for the asymptotic distribution of maximum score estimator is inconsistent. Their intuition carries through to our density-weighted MMS estimator in the partial identification scenario. Therefore, we propose to estimate the critical values by subsampling. Applying the results in Nolan and Pollard (1987, 1988), we show that the inferential statistic converges in distribution to a non-degenerate random variable, which ensures the validity of the subsampling procedure. In Section 4, we conduct Monte Carlo simulations under several choices of subsample sizes. The finite sample performance provides support to our inference procedure.

The key in our sample objective function is that it effectively controls the errors induced by the first stage nonparametric estimation in indicator functions. As in MT, our sample objective function also contains the term $\mathbf{1}\{\mathbb{E}(y|x, \nu_0, \nu_1) \geq 1-\alpha\}$ for some $\alpha \in (0, 1)$, which demands a nonparametric plug-in estimator of the conditional expectation inside the indicator function. By choosing bandwidths and kernels properly, we show that first stage estimation errors are asymptotically negligible and will not distort the asymptotic behavior of the second stage estimator.

Our method is also related to the literature of using $\mathscr{U}$-process theory to derive asymptotic properties of estimators, e.g. Sherman (1994b) establishes the asymptotic properties of the $\mathscr{U}$-process in the analysis of a generalized semiparametric regression model, which includes Ichimura (1993) and Klein and Spady (1993) as leading examples. The binary response model that we consider in this paper is different from Sherman (1994b) in two aspects. First, the parameters of interest are not point identified. Second, as a trade-off of the robustness from the conditional median assumption, our density-weighted MMS estimator has an "irregular" convergence rate which is slower than root-n. We do, however, discuss the extension of the density-weighting idea to a regular case—the parametric regression model with interval data. We propose a density-weighted modified minimum distance (MMD) method in a similar way to consistently estimate the identified set at a nearly parametric rate.

A line of literature on cube-root-n asymptotics has been developed for a variety of "irregular" estimators (see, e.g., Abrevaya, 2000). For the semiparametric binary response models, traditional maximum-score-type estimators have been reviewed, e.g. in Kim and Pollard (1990) and Horowitz (1998). The unusual cube-root-n convergence comes from the fact that maximum score sample criterion function is essentially a step function of parameters, which is "irregular" in the sense that it does not allow for a quadratic expansion.[2] Similar intuition carries through to the asymptotic analysis in our setting where the parameters of interest are partially identified: we show that the irregular set estimator converges to the identification region at a rate slightly slower than cube-root-n.[3] Blevins (2012) also studies the asymptotic problems of irregular set estimators, which is related to the present paper, but has a different focus.

The rest of the paper is organized as follows. Section 2 reviews the semiparametric response model with interval data and the MMS estimator proposed by MT. In Section 3, we introduce the density-weighted MMS estimator and provide the conditions for valid inference. Section 4 reports Monte Carlo experiment results. We discuss some possible extensions in Section 5 and conclude the paper in Section 6.

2. Semiparametric binary response model with interval data

Consider the following semiparametric binary response model studied in MT,

$$y = 1\left[x'\beta + \delta\nu + \epsilon > 0\right],$$

where $x \in \mathbb{R}^d$, $\nu \in \mathbb{R}$ and $\epsilon \in \mathbb{R}$. (y, x', ν_0, ν_1) are observed by researchers with $\nu_0 \leq \nu \leq \nu_1$. $\beta \in B \subset \mathbb{R}^d$ and $\delta \in \mathbb{R}$ are parameters of interest. The following assumption is made in MT, and throughout the present paper as well.

Assumption 2.1. Let Semiparametric Binary Regression (SBR) assumptions hold.

SBR-1 For a specified $\alpha \in (0, 1)$, $q_\alpha(\epsilon|x, \nu) = 0$. $\mathbb{P}(\epsilon \leq 0|x, \nu) = \alpha$.
SBR-2 $\mathbb{P}(\epsilon|x, \nu, \nu_0, \nu_1) = \mathbb{P}(\epsilon|x, \nu)$.
SBR-3 $\delta > 0$.

Assumption SBR-1 is the α-quantile-independence condition suggested by Manski (1975, 1985); SBR-2 asserts that observation of $[\nu_0, \nu_1]$ would not provide additional information for the distribution of ϵ if we know ν and x. SRB-2 holds if the bracket for each ν is generated at random, i.e., given x and ν, which bracket (with $\nu_0 \leq \nu \leq \nu_1$) to be reported has to be independent with ϵ (see Aucejo et al., 2013). In practice, if the set of brackets are predetermined for reporting ν and forms a partition on the real line, then the conditional distribution of ν_0 and ν_1 given ν is degenerate and SBR-2 holds trivially. Assumption SBR-3 is strong but could be substituted with weaker model restrictions that identify the sign of δ. In addition, positive δ constitutes a normalization.

As pointed out by MT, the threshold-crossing condition is invariant to the scale of the parameters. Hence, we set $\delta = 1$ throughout as a scale normalization. Further, MT characterize the sharp identification region of β by

$$B^* = \{b \in \mathbb{R}^k : \mathbb{P}[T(b)] = 0\}, \tag{1}$$

where $T(b) = \{(x, \nu_0, \nu_1) : (xb + \nu_1 \leq 0 < x\beta + \nu_0) \cup (x\beta + \nu_1 \leq 0 < xb + \nu_0)\}$.

MT propose a consistent set estimator for B^*: the modified maximum score (MMS) estimator. Let $z = (x', \nu_0, \nu_1)'$ and $P(z) = \mathbb{P}(y = 1|z)$. Let further $\lambda(z) = 1[P(z) > 1-\alpha]$ and $\mathrm{sgn}(\cdot)$ be the conventional sign function.[4]

2 Under additional smoothness assumptions on the error term's density, Horowitz (1992) propose a smoothed MSE, which has a limiting normal distribution and a rate of convergence that is at least $n^{-2/5}$ and can be arbitrarily close to $n^{-1/2}$.

3 On the other hand, a smoothed sample criterion function does not necessarily guarantee the corresponding estimator will converge at a parametric rate: in a simple setting of binary response models with a special regressor, Khan and Tamer (2010) show that the identification-at-infinity of parameters could also result in a convergence rate slower than the parametric rate. Chen et al. (2013) extend such a result.

4 We adopt the convention $sgn(0) = -1$.