ARTICLE IN PRESS

Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Econometrics

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

Varying coefficient panel data model in the presence of endogenous selectivity and fixed effects☆

Emir Malikov [a,*], Subal C. Kumbhakar [b], Yiguo Sun [c]

[a] Department of Economics, St. Lawrence University, Canton, NY, United States

[b] Department of Economics, State University of New York at Binghamton, Binghamton, NY, United States

[c] Department of Economics and Finance, University of Guelph, Guelph, ON, Canada

ARTICLE INFO

Article history:
Available online xxxx

JEL classification:
C14
C33
C34
G21

Keywords:
Credit union
Fixed effects
Selection
Semiparametric
Smooth coefficient
Switching regression
Varying coefficient

ABSTRACT

This paper considers a flexible panel data sample selection model in which (i) the outcome equation is permitted to take a semiparametric, varying coefficient form to capture potential parameter heterogeneity in the relationship of interest, (ii) both the outcome and (parametric) selection equations contain unobserved fixed effects and (iii) selection is generalized to a polychotomous case. We propose a two-stage estimator. Given consistent parameter estimates from the selection equation obtained in the first stage, we estimate the semiparametric outcome equation using data for the observed individuals whose likelihood of being selected into the sample stays approximately the same over time. The selection bias term is then "asymptotically" removed from the equation along with fixed effects using kernel-based weights. The proposed estimator is consistent and asymptotically normal. We showcase our estimator by applying it to study production technologies of US retail credit unions from 2002 to 2006.

1. Introduction

Semiparametric methods have become a part of a standard methodological toolkit of applied researchers in economics. These methods are attractive for their ability to circumvent limitations of conventional parametric models by allowing more flexible specifications and thus mitigating (at least partly) the risk of misspecification. While they admittedly require more prior assumptions and therefore are not as flexible as their (completely) nonparametric counterparts, semiparametric models have nevertheless gained popularity due to their capability to alleviate the so-called "curse of dimensionality" associated with nonparametric estimation.

This paper considers a particular class of semiparametric models in which parameters of a linear regression are permitted to be unspecified smooth functions of some variables (Hastie and Tibshirani, 1993; Cai et al., 2000; Li et al., 2002). Such "varying coefficient" (hereinafter VC) models[1] have recently become a subject of prolific research in the econometric literature that attempts to extend the method to new settings. For instance, Das (2005), Cai et al. (2006) and Cai and Xiong (2012) consider VC models in the presence of endogenous variables and propose applying instrumental variables approach to tackle the endogeneity problem. However, the overwhelming majority of these studies place the model either in the cross-sectional (as in the above cited papers) or in the time series settings [e.g., Cai (2007)]. Analysis of VC models in a panel data setting is however relatively scarce, arguably due to difficulties associated with tackling unobserved effects. For instance, Cai and Li (2008) study a VC model in the dynamic panel setting that assumes any unobserved effects away. Sun et al. (2009) somewhat fill the void by proposing a VC panel data model estimator which allows treatment of both random and fixed effects.[2]

☆ We would like to thank Peter Schmidt, referees and conference participants at the 2014 NY Camp Econometrics IX for many helpful comments and suggestions. Any remaining errors are our own.

* Corresponding author.
E-mail addresses: emalikov@stlawu.edu (E. Malikov), kkar@binghamton.edu (S.C. Kumbhakar), yisun@uoguelph.ca (Y. Sun).

http://dx.doi.org/10.1016/j.jeconom.2015.06.007
0304-4076/

[1] Such models are also referred to as "smooth coefficient" or "functional coefficient" models.

[2] The studies of nonparametric panel data models that consider the presence of either random or fixed effects include, e.g., Das (2003), Henderson and Ullah (2005) and Henderson et al. (2008). Alternatively, there are studies that focus on panel

However, the semiparametric literature has broadly overlooked another important feature of the data that applied researchers often have to deal with: namely, the presence of selectivity. Such a problem is acute in studies of wage and labor supply decisions that go back to Heckman's (1974,1979) seminal work and many other labor economics applications and not only. In this paper, we therefore take a semiparametric VC model a step further by considering it in the panel data setting and the presence of endogenous selection and fixed effects.[3] For a similar model in a cross-sectional setting, see Das et al. (2003), whose model allows both the outcome and selection equations to take completely nonparametric forms. Das (2004) extends the above model to a panel data case with (exogenous) random effects.

Thus, we consider a flexible panel data sample selection model in which (i) the outcome equation is permitted to take a semiparametric VC form to capture potential parameter heterogeneity in the relationship of interest, (ii) both the outcome and selection equations contain unobserved fixed effects and (iii) selection is generalized to a polychotomous case. In this paper, we restrict our analysis to models with parametric selection equations. Our model can be considered as a generalization of conventional parametric panel data sample selection models [see Baltagi (2013) for a comprehensive review]. Relatively few such parametric models allow for a fixed-effect type heterogeneity. For instance, in the case of strictly exogenous right-hand-side covariates, Wooldridge (1995), Rochina-Barrachina (1999) and Malikov and Kumbhakar (2014) propose correlated effects estimators, whereas Kyriazidou (1997) develops an estimator that allows for completely unspecified fixed effects in both the selection and outcome equations. In this paper, we let fixed effects to be correlated with the right-hand-side covariates in an arbitrary way and remove them "nonparametrically", which makes Kyriazidou's (1997) estimator be the closest parametric counterpart to the semiparametric one that we propose in this paper. The difference between the two lies in the facts that we let the outcome equation take a more flexible VC form and that we generalize selection to a polychotomous case.

We propose estimating our model in two stages. We suggest consistently estimating the selection equation in the first stage via any of several methods available in the literature such as Manski's (1987) and Horowitz's (1992) (smoothed) conditional maximum score or Chamberlain's (1980) conditional logit estimators. The obtained estimates can then be used to evaluate the conditional probability of an individual to be selected into the sample in each time period. In the second stage, we propose estimating the VC outcome equation using data for observed individuals (cross-sections) whose estimated likelihood of being selected into the sample stays approximately the same over time. For such individuals, the sample selection bias would be approximately time-invariant and thus can be treated as another component of fixed effects present in the outcome equation. Given that there are unlikely to be many (if any at all) cross-sections with exactly the same selection probabilities over time, we adopt the idea of Ahn and Powell (1993) and Kyriazidou (1997) and weigh these cross-sections based on "closeness" of their respective selection probabilities (and thus their selectivity biases) to being the same over time. The weighted semiparametric outcome equation can then be estimated in a manner similar to that proposed by Sun et al. (2009). The selection bias term is "asymptotically" removed from the equation along with fixed effects using kernel-based weights. The latter is advantageous over conventional first-differencing[4] because it mitigates the need to use backfitting and allows identification of an intercept coefficient function. We show that, under appropriate assumptions on the rate of convergence of the first-stage estimator of the selection equation, our proposed estimator is consistent and asymptotically normal.

We first investigate the finite sample performance of the proposed estimator in a small Monte Carlo simulation. The results are encouraging and show that, in the presence of endogenous selectivity, our estimator is less biased than a "naive" estimator which overlooks the selection issue. We also find that the estimation becomes more stable as the sample size increases. To conserve space, we relegate the detailed discussion of a Monte Carlo study to online supplement to the paper (see Appendix C).

We next apply our estimator to study production technologies of US retail credit unions in the period from 2002 to 2006. There has recently been a substantial interest in investigation of credit unions' production technologies, given a dramatic transformation that the US credit union industry has been undergoing over the past few decades.[5] Copious mergers and acquisitions have transformed the industry from one which had primarily consisted of small-scale local institutions catering to a handful of members to a now trillion dollar industry that constitutes a significant portion of the US financial services markets, serving a hundred million customers in the country [authors' calculations based on National Credit Union Administration (2011)].

Studies that have investigated the performance of US credit unions had to deal with the problem of having a large number of observations for which the reported values of credit unions' outputs are zeros. Researchers have handled this problem either by linearly aggregating all types of outputs into a single bundle [e.g., Fried et al. (1999), Wheelock and Wilson (2011) and Wheelock and Wilson (2013)] or by replacing zero outputs with an arbitrarily chosen small positive number (Frame et al., 2003). The presence of zero-value observations is however likely to be informative and may indicate significant differences among credit unions in terms of the service menu they offer to members. Ignoring this *observed* heterogeneity in the provision of services amounts to making a strong and rather unrealistic assumption that all credit unions share the same "production" technology that is invariant to the menu of services they provide. This assumption of homogeneous technology across credit unions is likely to result in the loss of information and the misspecification of the econometric model, which is further aggravated if the choice of the differing service menus by credit unions is endogenous (Malikov et al., 2014). In this paper, we model this observed heterogeneity as an outcome of an endogenous choice (selection). Moreover, we also allow for *unobserved* heterogeneity among credit unions, something that has been broadly overlooked in most existing studies.

We find some significant distortions in cost elasticity estimates if one ignores selectivity. Similarly, we document dramatic differences in elasticity estimates between our VC sample selection model and its parametric counterpart. We find that the estimated relationship between scale economies and the smoothing variable (here, the asset size) from our VC model is quite different from that implicitly implied by a parametric model. These findings call for extra caution when researchers first estimate a parametric model of credit union production technologies (even after controlling

data applications of other classes of semiparametric models such as Li and Stengos (1996), Su and Ullah (2006) and Lin and Carroll (2006).

[3] Here, we focus on a panel data application, given its increasing availability to researchers (as opposed to mere cross-sectional data). We do not consider the case of random effects because applied researchers often consider the assumption of exogenous heterogeneity unsupported by the data and difficult to justify. The violation of such an assumption would yield inconsistent estimates.

[4] For instance, Kyriazidou (1997) proposes applying first-differencing in order to purge the sample selection term and fixed effect from the outcome equation.

[5] See Wheelock and Wilson (2011), Malikov et al. (2014) and references therein.

Please cite this article in press as: Malikov, E., et al., Varying coefficient panel data model in the presence of endogenous selectivity and fixed effects. Journal of Econometrics (2015), http://dx.doi.org/10.1016/j.jeconom.2015.06.007