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

Contents lists available at ScienceDirect

Journal of Econometrics

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

Nonparametric identification and estimation of sample selection models under symmetry

Songnian Chen [a,*], Yahong Zhou [b], Yuanyuan Ji [b,c]

[a] *Hong Kong University of Science and Technology, Hong Kong*
[b] *Shanghai University of Finance and Economics, China*
[c] *Shanghai Academy of Social Sciences, China*

ARTICLE INFO

Article history:
Received 27 October 2016
Received in revised form 5 June 2017
Accepted 27 September 2017
Available online xxxx

JEL classification:
C13
C14

Keywords:
Sample selection
Nonparametric estimation
Symmetry

ABSTRACT

Under a conditional mean restriction Das et al. (2003) considered nonparametric estimation of sample selection models. However, their method can only identify the outcome regression function up to a constant. In this paper we strengthen the conditional mean restriction to a symmetry restriction under which selection biases due to selection on unobservables can be eliminated through proper matching of propensity scores; consequently we are able to identify and obtain consistent estimators for the average treatment effects and the structural regression functions. The results from a simulation study suggest that our estimators perform satisfactorily.

1. Introduction

Since Heckman's (1974) seminal work, sample selection models have been widely used in applied research in correcting for bias arising from non-random sampling, which includes applications in modeling the impact of unions, occupational choice, the choice of region of residence and choice of industry, among others. Heckman's two-step estimator and its extension to semiparametric two-step estimators of Newey (1988), Powell (1989), Ahn and Powell (1993), Chen (1999) and Chen and Zhou (2010) require parametric specification for either the regression function or the error distribution, or both. Therefore, these parametric and semiparametric two-step estimators are, in general, not robust to misspecification of the functional form of the regression function or the parametric error distribution.

Das et al. (2003) considered two-step nonparametric estimation of sample selection models under a conditional mean restriction that allows for the same degree of flexibility as standard nonparametric regression. However, in their analysis, the regression function in the outcome equation is only identified up to an unknown constant, and as a result, their method cannot be used to identify or estimate either conditional or unconditional average treatment effects. In this paper, we also consider nonparametric identification and estimation of the sample selection model, but strengthen the conditional mean restriction in Das et al. (2003) to a joint symmetry restriction. Under this shape restriction on the error distribution, selection biases due to selection on unobservables can be eliminated through proper matching of propensity scores and as a result, we are able to identify and construct consistent estimators for the conditional and unconditional average treatment effects, as well as the structural regression function in the outcome equations. Under the symmetry restriction, we show that selection bias can be eliminated through an appropriate weighting scheme, leading to nonparametric identification and consistent estimation of the outcome regression function and the average treatment effects.

Symmetry restrictions on underlying error distribution have been widely used in the literature. In the program evaluation literature, various alternative assumptions have been imposed for the identification of the average treatment effect (ATE). For the control function approach (e.g., Heckman and Navarro-Lozano, 2004), the idea of "Identification at infinity" (Heckman, 1990), which requires a "large support" condition, has played an important role in identifying the average treatment effect (ATE). As pointed out by Heckman and Navarro-Lozano (2004), in practice, the "large support" condition is often unrealistic, and consequently they imposed the joint symmetry restriction instead of appealing to the "identification at infinity" for the identification of "ATE". In the context of binary discrete outcomes, Aakvik et al. (1999), partly

---

[*] Correspondence to: Department of Economics, Hong Kong University of Science and Technology, Clearwater Bay, Kowloon, Hong Kong.
E-mail address: snchen@ust.hk (S. Chen).

https://doi.org/10.1016/j.jeconom.2017.09.004

Please cite this article in press as: Chen S., et al., Nonparametric identification and estimation of sample selection models under symmetry. Journal of Econometrics (2017), https://doi.org/10.1016/j.jeconom.2017.09.004.

motivated by Chen (1999), also exploited the joint symmetry restriction for identification of average treatment effect (ATE).

From a different perspective, Angrist (2004) considered various assumptions under which IV estimates have broader predictive power beyond the compliers group. In particular, he is interested in the assumptions that link a local Average Treatment Effect (LATE) to the population Average Treatment Effect (ATE), which is not instrument-dependent. Among various assumptions considered, Angrist (2004) noted that the symmetry assumption is more appealing because it is not fundamentally inconsistent with the benchmark Roy-type selection model, unlike "no selection bias" or "conditional constant effects" assumptions. Angrist (2004) suggested that intuitively, with symmetrically distributed latent errors in the index framework, together with a symmetric first stage, the LATE becomes equivalent to the ATE because average treatment effects for individuals with characteristics of the compliers are representative of average treatment effects for individual over the entire distribution.[1] Angrist (2004) illustrated these ideas using sibling-sex composition to estimate the effect of childbearing on economics and marital outcomes; in particular, Angirst found in the study that for teen mothers, LATE is indeed identical to the population average treatment effect ATE when the latter is imputed under joint symmetry restriction.

While Aakvik et al. (1999) and Angrist (2004) exploited the joint symmetry restriction to identify the average treatment effect (ATE), Chen and Khan (2010) made use of the joint symmetry restriction, together with an "equality" condition,[2] to identify the average treatment effect (ATE) on wage inequality. Using the approach developed by Chen and Khan (2010), Antonczyk (2011) estimated the ATE of collective bargaining on the dispersion of wages in Germany.

Symmetry restrictions on underlying error distributions have also been used in other models. Chen (2000) and Chen et al. (2016) studied the identification and estimation of binary choice models under the symmetry restriction. Powell (1986), Honoré (1992) and Dong and Lewbel (2011) are yet more models that exploit the symmetry restriction.

The paper is organized as follows. Section 2 presents the model and discusses the identification and estimation issues. Section 3 contains the large sample results of the proposed estimators. We provide the results of a Monte Carlo simulation study in Section 4. Section 5 contains the concluding remarks. The proof of the main theorem is relegated to the Appendix.

2. The model and estimator

We consider the nonparametric switching regression model

$$y_i = y_{1i}d_i + y_{0i}(1-d_i) \qquad (i = 1, \ldots, n) \tag{2.1}$$

where

$$y_{1i} = g_1(x_i) + u_{1i} \tag{2.2}$$

and

$$y_{0i} = g_0(x_i) + u_{0i} \tag{2.3}$$

denote the outcome equations under regimes 0 and 1, and the selection equation is of the form

$$d_i = 1\left\{m(w_i) > v_i\right\}. \tag{2.4}$$

Standard sample selection models correspond to the special case where y_{0i} is identically zero. Here $g_1(x)$ and $g_0(x)$ are unknown functions in the outcome equations, d is the binary selection indicator, $m(w)$ is an unknown function, $x \in R^{d_x}$ and $w \in R^{d_w}$ are vectors of the regressors with possibly overlapping components, and (u_1, u_0, v) are the unobservable error terms independent of (x, w) with $E(u_0|x, w) = 0$. As in Das et al. (2003), we also impose an exclusion restriction such that w contains some component not in x.

The main equation can be written as a nonparametric regression model with a random coefficient

$$y_i = g_0(x_i) + d_i(\alpha(x_i) + \varepsilon_i) + u_{0i} \tag{2.5}$$

where $\alpha(x_i) = g_1(x_i) - g_0(x_i)$ and $\varepsilon_i = u_{1i} - u_{0i}$. In this paper we consider the identification and estimation of $\alpha(x)$ and $\alpha = E[\alpha(x_i)]$, or $\alpha_{\mathcal{S}} = E_{\mathcal{S}}[\alpha(x_i)] = E[\alpha(x_i)|x_i \in \mathcal{S}]$, for some fixed set,[3] $\mathcal{S}$ under the condition that (ε_i, v_i) is independent of (x_i, w_i) and symmetrically distributed around the origin, and $E(u_0|x, w) = 0$. Also, assume $P(w) = \Pr(m(w_i) > v_i|w_i = w) = F_v(m(w))$ is a strictly increasing function of $m(w)$.

Under the conditional mean restriction that $E(u_0|x, w) = E(u_1|x, w) = 0$, Das et al. (2003) considered identification and estimation of the nonparametric sample selection model. However, their approach only provides the identification result and consistent estimators for $g_0(x)$ and $g_1(x)$ up to unknown constant terms; as a result, the conditional and the unconditional average treatment effects $\alpha(x)$ and $\alpha_{\mathcal{S}}$ are not identified in their context. In contrast, under the joint symmetry restriction, we are able to identify and consistently estimate $\alpha(x)$ and $\alpha_{\mathcal{S}}$ (similarly, $g_0(x)$ and $g_1(x)$ as well). Under the conditional mean restriction with a special regressor, Lewbel (2007) considered the estimation of a more general model which includes the switching regression model as a special case; however, Lewbel (2007) requires a large support condition for the special regressor, which is not needed for our approach here.

To motivate our identification and estimation approach, consider the regression function

$$\begin{aligned} g(x, P) &= E(y_i|x_i = x, P(w_i) = P) \\ &= g_0(x) + P\alpha(x) + E(d_i\varepsilon_i|x_i = x, P(w_i) = P) \\ &= g_0(x) + P\alpha(x) + \lambda(P) \end{aligned} \tag{2.6}$$

where the selection bias term $\lambda(P)$, which only depends on the propensity score P in this index sufficiency framework, Heckman et al. (1998) in general, does not vanish. However, under the index sufficiency and joint symmetry, the selection bias term $\lambda(P)$ can be shown to be symmetric around $P = 1/2$. This symmetry result is exploited below to achieve the cancellation of the selection bias terms and thus facilitate the identification and estimation of the treatment effects parameters.

Proposition 1. *If (ε_i, v_i) is independent of (x_i, w_i) and symmetrically distributed around the origin, $P(w) = F_v(m(w))$ is a strictly increasing function of $m(w)$, then the selection bias term $\lambda(P)$ is symmetric around $P = 1/2$; namely, $\lambda(P) = \lambda(1-P)$.*

Proof. Under the independence and symmetry condition, and the fact that $F_v(\cdot)$ is a strictly increasing function, it is easy to see

$$F_v^{-1}(P) + F_v^{-1}(1-P) = 0.$$

---

[1] Alternatively, it implies that expected outcomes for compliers can be obtained as the average of expected outcomes for always and never-takers.

[2] In a sense, the "equality" condition resembles the "similarity" condition of Chernozhukov and Hansen (2005).

[3] Here $\mathcal{S}$ is the set of the common support of x_i for the two regimes, or a subset of it; see Heckman et al. (1998) for some detailed discussions.

Please cite this article in press as: Chen S., et al., Nonparametric identification and estimation of sample selection models under symmetry. Journal of Econometrics (2017), https://doi.org/10.1016/j.jeconom.2017.09.004.