Journal of Econometrics 181 (2014) 77–91

Contents lists available at ScienceDirect

Journal of Econometrics

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

A flexible parametric approach for estimating switching regime models and treatment effect parameters[☆]

Heng Chen[a], Yanqin Fan[b,*], Jisong Wu[c]

[a] Bank of Canada, ON, Canada
[b] University of Washington, WA, USA
[c] Indiana University-Purdue University Indianapolis, IN, USA

ARTICLE INFO

Article history:
Received 11 August 2009
Received in revised form
3 April 2013
Accepted 27 June 2013
Available online 10 May 2014

JEL classification:
C31
C34
C51

Keywords:
EM algorithm
Generalized normal mean–variance mixture
Inverse-mills ratio
Skewness
Two-step estimation

ABSTRACT

In this paper, we propose a flexible, parametric class of switching regime models allowing for both skewed and fat-tailed outcome and selection errors. Specifically, we model the joint distribution of each outcome error and the selection error via a newly constructed class of multivariate distributions which we call generalized normal mean–variance mixture distributions. We extend Heckman's two-step estimation procedure for the Gaussian switching regime model to the new class of models. When the distributions of the outcome errors are asymmetric, we show that an additional correction term accounting for skewness in the outcome error distribution (besides the analogue of the well known inverse mill's ratio) needs to be included in the second step regression. We use the two-step estimators of parameters in the model to construct simple estimators of average treatment effects and establish their asymptotic properties. Simulation results confirm the importance of accounting for skewness in the outcome errors in estimating both model parameters and the average treatment effect and the treatment effect for the treated.

Published by Elsevier B.V.

1. Introduction[1]

Switching regime models (SRMs) extend the Roy model of self-selection by allowing a more general decision rule for selecting into different states. The income maximizing Roy model of self-selection was developed to explain occupational choice and its consequences for the distribution of earnings when individuals differ in their endowments of occupation-specific skills, see Heckman and Honore (1990). By allowing a more general decision/selection rule, SRMs enjoy a much wider scope of applications than the Roy model. Recently, SRMs have been used to evaluate average effects of a policy intervention using choice data. Heckman et al. (2003) derive expressions for four treatment effect parameters for a Gaussian copula SRM and a Student's t copula SRM with normal outcome errors and non-normal selection error.[2] These are the average treatment effect (ATE), the treatment effect for the treated (TT), the local average treatment effect (LATE, Imbens and Angrist, 1994), and the marginal treatment effect (MTE, Bjorklund and Moffitt, 1987; Heckman, 1997; Heckman and Vytlacil, 1999, 2000a,b, 2005).

One of the most commonly used SRMs in empirical work is the Gaussian SRM in which the vector of outcome errors and the selection error follows a trivariate normal distribution. One reason for its popularity is the simplicity of Heckman's two-step procedure for estimating the Gaussian SRM introduced in Heckman (1976). The Gaussian SRM has been extended to allow for non-normal marginal distributions in the errors in Lee (1982, 1983),

☆ This is a substantially revised version of a previous paper entitled "Simple Estimators of Switching Regimes Models with Normal Mean-Variance Mixture Copulas and Average Treatment Effects". We are grateful to the Editor, Cheng Hsiao, and two anonymous referees for their insightful comments which have led to the current much improved paper.

* Correspondence to: Department of Economics, University of Washington, Seattle, WA, 98195-3330, USA. Tel.: +1 206 543 8172.
E-mail address: fany88@u.washington.edu (Y. Fan).

1 Parts of this section are based on the insightful comments of an anonymous referee on an earlier version of this paper.

2 They did not use the concept of copulas, but their models can be interpreted in this way.

http://dx.doi.org/10.1016/j.jeconom.2013.06.006
0304-4076/Published by Elsevier B.V.

Heckman et al. (2003), and Li et al. (2004). The models in Heckman et al. (2003) essentially assume that the trivariate error vector follows a distribution with either the Gaussian copula or a trivariate Student's t copula. When the outcome errors are normal or Student's t and the selection error has any parametric distribution (known up to a finite dimensional parameter vector), Lee (1982, 1983), and Heckman et al. (2003) show that the model parameters can be consistently estimated by a two-step procedure extending Heckman's two-step procedure for the Gaussian SRM. By extending Heckman's two-step procedure to Student's t outcome error(s), fat-tailed outcome error(s) can be accounted for in the two-step procedure of Lee (1982, 1983), Heckman et al. (2003). Indeed simulation results in Heckman et al. (2003) show that their two-step estimation procedure works well for fat-tailed distributions when used to estimate ATE, TT, and LATE and that for ATE, TT, even Heckman's two-step procedure based on the Gaussian SRM yields minor biases. For skewed outcome distributions, however, the results are completely different; the sensitivity of parameter estimates in the Gaussian SRM to asymmetric outcome distributions was long recognized in the statistics literature, see Little (1982) and references therein. Although not reported in their paper, Heckman et al. (2003) mention in their Footnote 7 that 'When generating data from highly asymmetric distributions, such as a χ^2 (3), we do see larger biases' (in the estimates of ATE and TT).

Work in the semiparametric literature, on the other hand, estimate SRMs without imposing parametric distributional assumptions on the error vector, see, e.g., Newey et al. (1990), Ahn and Powell (1993) and Das et al. (2003), among others. These estimates are thus robust to possible misspecification in the distribution of the error vector. However, as pointed out in Heckman (1990), most of these approaches only estimate slope parameters of the outcome equations excluding intercepts of the outcome equations, and thus cannot be used to estimate average treatment effect parameters. Furthermore, without additional distributional assumptions, it is not possible to identify the standard treatment effect parameters without invoking identification-at-infinity arguments that require instruments with unbounded support, and even with such assumptions, the standard treatment effect parameters can only be estimated at a nonparametric rate, see e.g., Andrews and Schafgans (1998). The recent work by Heckman and Vytlacil (1999, 2000a,b, 2005) and others have been semiparametric and focused on average treatment effect parameters, but again they face the trade-off between pursuing nontraditional treatment effect parameters (the MTE parameter) for which they can only obtain a slow rate of convergence, or the traditional treatment effect parameters for which they either need to do a bounding analysis or have large support assumptions in which case they identify the standard treatment effect parameters that can only be estimated at a slow rate of convergence. The requirement for extremely large samples to implement the Heckman–Vytlacil approach is one of the reasons to consider parametric estimation as in Heckman et al. (2003). Likewise, the inability of these models to identify traditional treatment effect parameters without relying on identification-at-infinity is the reason that Chamberlain (2010) advocates imposing a flexible parametric model.

Chen (1999) presents an alternative semiparametric approach to identify the intercepts of the outcome equations and estimate them at the parametric rate. But his approach relies critically on symmetry of the distribution of the error vector. This paper proposes a flexible, parametric method for estimating SRMs and the corresponding treatment effect parameters allowing for asymmetric and fat-tailed outcome (and selection) errors. It makes two main contributions to the econometrics and statistics literatures. First, it introduces the class of multivariate Generalized Normal Mean–Variance Mixture (GNMVM) distributions and the associated copulas. To the best of the authors' knowledge, the class of multivariate GNMVM distributions is new. It overcomes the severe drawback of the class of multivariate NMVM distributions that the only multivariate NMVM distribution with independent marginal distributions is that of multivariate normal, see McNeil et al. (2005). In contrast, the class of multivariate GNMVM allows for independent, but skewed/non-normal marginal distributions. Copulas associated with GNMVM distributions include Gaussian, Student's t, and skewed t copulas. As shown in Demarta and McNeil (2005), skewed t copulas exhibit asymmetric and different left and right tail dependence. Using the class of multivariate GNMVM distributions/copulas, we construct a new class of SRMs referred to as the GNMVM-SRMs/GNMVMC-SRMs in which the bivariate distributions of each outcome error and the selection error are assumed to follow either a GNMVM distribution or a distribution with a GNMVM copula and any parametric marginal distributions. The Gaussian copula and Student's t copula SRMs in Heckman et al. (2003) are members of GNMVMC-SRMs.[3] Gaussian and Student's t SRMs in Chib (2005) are members of GNMVM-SRMs which also include SRMs with multi-modal and skewed error distributions. More importantly, the class of GNMVM-SRMs allows for independent asymmetric outcome errors and selection error, so asymmetry in the observed outcome distributions could be due to either asymmetric outcome errors or selection.

The second contribution of this paper is to develop a two-step estimation procedure for the class of GNMVM-SRMs further extending Heckman's two-step procedure for the Gaussian SRM. In the first step, we propose a novel simulation-based EM algorithm for estimating the selection equation and in the second step, we use OLS to estimate two linear regressions each with two correction terms resulting in estimators of parameters in each outcome equation. Using the two-step procedure, we construct estimators of ATE, TT, LATE, and MTE in GNMVM-SRMs and establish their asymptotic properties. In addition to the extension of the Inverse-Mills ratio in Heckman's two-step procedure, there is an additional correction term in the second step regression for estimating GNMVM-SRMs. As a result, applying Heckman's two-step procedure to SRMs with skewed outcome distributions may lead to inconsistent estimators of parameters in the potential outcome equations and of the four treatment effect parameters. Our simulation results using asymmetric outcome distributions also reveal large biases in the estimates of parameters in SRMs and/or of ATE and TT if skewness is not accounted for in the estimation procedure. In contrast, our two-step procedure allowing for skewness in the outcome distributions performs very well. We also extend our two-step estimation procedure for GNMVM-SRMs to the subclass of GNMVMC-SRMs in which the outcome errors follow univariate NMVM distributions and the selection error follows any parametric distribution. In general, the second step for the subclass of GNMVMC-SRMs may involve a nonlinear regression, but for certain members of GNMVM copulas such as Gaussian copula or Student's t copula with a known degree of freedom in Lee (1982, 1983), Heckman et al. (2003), the second step regression is linear.

The rest of this paper is organized as follows. In Section 2, we introduce the class of GNMVMC-SRMs and some special cases. In Section 3, we propose a two-step procedure for estimating parameters in the potential outcome equations in GNMVM-SRMs and GNMVMC-SRMs when the outcome errors follow NMVM distributions. In Section 4, we use our two-step estimation procedure to construct estimators of ATE, TT, LATE, and MTE, extending the estimators of Heckman et al. (2003) to a much wider class of SRMs. We present results from a small Monte Carlo simulation study in Section 5. Section 6 concludes. Technical proofs are relegated to the Appendix.

[3] In fact, Heckman et al. (2003) impose a trivariate Gaussian or Student's t copula structure on the trivariate error vector as opposed to bivariate Gaussian or Student's t copula structure on each bivariate vector of an outcome error and selection error separately.