Journal of Econometrics 199 (2017) 63–73

Contents lists available at ScienceDirect

Journal of Econometrics

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

Identification in a generalization of bivariate probit models with dummy endogenous regressors[☆]

Sukjin Han[a,*], Edward J. Vytlacil[b]

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

ARTICLE INFO

Article history:
Received 4 September 2013
Received in revised form 25 December 2016
Accepted 1 April 2017
Available online 27 April 2017

Keywords:
Identification
Triangular threshold crossing model
Bivariate probit model
Dummy endogenous regressors
Binary response
Copula
Exclusion restriction

ABSTRACT

This paper provides identification results for a class of models specified by a triangular system of two equations with binary endogenous variables. The joint distribution of the latent error terms is specified through a parametric copula structure that satisfies a particular dependence ordering, while the marginal distributions are allowed to be arbitrary but known. This class of models is broad and includes bivariate probit models as a special case. The paper demonstrates that having an exclusion restriction is necessary and sufficient for global identification in a model without common exogenous covariates, where the excluded variable is allowed to be binary. Having an exclusion restriction is sufficient in models with common exogenous covariates that are present in both equations. The paper then extends the identification analyses to a model where the marginal distributions of the error terms are unknown.

1. Introduction

This paper examines the identification of a class of bivariate threshold crossing models that nests bivariate probit models as a special case. The bivariate probit model was introduced in Heckman (1978) as one specification of simultaneous equations models for latent variables, and is commonly used in applied studies, such as Evans and Schwab (1995), Neal (1997), Goldman et al. (2001), Altonji et al. (2005), Bhattacharya et al. (2006), and Rhine et al. (2006), to name a few. Although the model has drawn much attention in the literature, relatively little research has been done to analyze the identification even in this restricted model.[1]

There are three papers in the literature that have studied identification of bivariate probit models: Freedman and Sekhon (2010), Wilde (2000), and Meango and Mourifié (2014). Freedman and Sekhon (2010) provide formal identification results for bivariate probit models, though they assume (and their proof strategy critically relies upon the assumption) that one of the exogenous regressors has large support. The large support condition is restrictive and limits the applicability of their analysis. Wilde (2000) also considers the identification of bivariate probit models. His identification analysis is limited to simply counting the number of unknown parameters and number of informative non-redundant probabilities in the likelihood function, i.e., the number of equations. His analysis only establishes a necessary condition for global identification since there may still exist multiple solutions in a system of nonlinear equations where the number of equations is at least as large as the number of unknown parameters. In fact, Meango and Mourifié (2014) show that, using as many equations as the number of parameters, there can be multiple solutions in a bivariate probit model where there are common *binary* exogenous regressors but no excluded instruments.[2]

In this paper, we derive identification results for a class of models specified by a triangular system of two equations with binary endogenous variables, where we generalize the bivariate normality assumption on the latent error terms of a bivariate probit model through the use of copulas. In particular, instead of requiring that the joint distribution of latent error terms be bivariate normal, we allow the marginal distributions to be arbitrary but known,

☆ We thank the coeditor, the associate editor and three anonymous referees for useful comments and suggestions.

* Corresponding author.
E-mail addresses: sukjin.han@austin.utexas.edu (S. Han), edward.vytlacil@yale.edu (E.J. Vytlacil).

[1] Heckman (1978) discusses identification via a maximum likelihood estimation framework in a model where one of the latent dependent variables is observed in the simultaneous equations model. In a framework where both are not observed, however, identification analysis through calculating the second derivative of a maximum likelihood criterion function is problematic since it is analytically hard to solve.

[2] Building upon Meango and Mourifié (2014) and the present paper, Han and Lee (2017) show that the solution is not unique even when exploiting the *full* set of equations implied by the model. These results demonstrate that Wilde's (2000) counting exercise is not sufficient for identification analysis.

http://dx.doi.org/10.1016/j.jeconom.2017.04.001
0304-4076/

while restricting their dependence structure by imposing that their copula function belongs to a broad class of parametric copulas that includes the normal copula as a special case. We then extend the results to a model where the marginal distributions are unknown. All results derived in this paper also apply to the special parametric case of bivariate probit models.

We first provide identification results in a model without common exogenous regressors, showing that, in such a model, having a valid exclusion restriction (i.e., instrument) is necessary and sufficient for global identification of the model. Unlike Freedman and Sekhon (2010), this result does not require a full support condition, and holds even if the instrument is binary. While Wilde (2000) restricts his analysis to bivariate probit models, we show that a bivariate normal distribution is not necessary for our identification strategy to work as long as a certain dependence structure is maintained. We extend the result to allow for the possibility of exogenous covariates that enter both equations and the possibility of instruments Z being vector valued without requiring any element of Z to be binary. Having an exclusion restriction is sufficient for identification in this context.[3] In this full model, we also provide identification results without assuming that the marginal distributions of the error terms are known. The structural parameters are shown to be identified under similar conditions as in the known-marginal case and the marginal distributions are shown to be additionally identified under a stronger support condition.

We make use of copulas to characterize the joint distribution of the latent error terms, which allows us to separate the error terms' dependence structure from their marginal distributions. Our analysis shows that identification is obtained through a condition on the copula, with the shape of the marginal distributions playing no role in the analysis. The condition we impose on the copula is that it satisfies a particular dependence ordering with respect to a single dependence parameter. Specifically, the condition is that the copula is ordered by a dependence parameter that is informative about the degree of dependence in the sense of the first-order stochastic dominance "FOSD". We show that this condition is satisfied by a broad range of single-parameter copulas including the normal copula. Thus, the assumption used in the literature that the latent variables follow a bivariate normal distribution is not critical in deriving identification results in this type of models.[4] We also introduce a novel dependence ordering concept that characterizes minimal structure on the copula that is required for our identification results. This ordering is more general than the FOSD ordering but slightly less interpretable.

Our use of copulas is related to Lee (1983), who uses a normal copula to generalize normal selection models. Chiburis (2010) is also related to our analysis. He introduces a normal copula to characterize the joint distribution of latent variables in a similar setting as in this paper, although no rigorous identification analysis is conducted for our class of models. To facilitate their inference procedure in a censored linear quantile regression model, Fan and Liu (2015) introduce one-parameter ordered families of Archimedean copulas in characterizing dependence between the dependent variable and censoring variable, but the ordering concept which defines their class of copulas differs from ours. Copulas have also been used to model the joint distributions of error terms in switching regime models (Fan and Wu, 2010) or the joint distribution of potential outcomes in randomized experiment settings (Fan and Park, 2010), where bounds on the distribution of treatment effects are derived. There are also recent papers that generalize a bivariate probit model using a copula structure (Winkelmann, 2012) or using nonparametric index functions instead of linear functions (Marra and Radice, 2011), or both (Radice et al., 2015), but all of these papers rely on the counting exercise for identification analysis.

The paper is organized as follows. In Section 2, we introduce the model and preliminary assumptions. Section 3 introduces dependence orderings and related concepts that are used to define the class of models we analyze. Section 4 shows identification of a simple, special case of our model, which is useful for subsequent analyses. Section 5 extends the identification analysis to the full model. Section 6 extends the results of the previous section to the case of nonparametric marginal distributions. Section 7 concludes with discussions on estimation and inference.

2. The model

Let Y denote the binary outcome variable and D the observed binary endogenous treatment variable. Let $\underset{(k+1)\times 1}{X} \equiv (1, X_1, \ldots, X_k)'$ denote the vector of regressors that determine both Y and D, and let $\underset{l\times 1}{Z} \equiv (Z_1, \ldots, Z_l)'$ denote a vector of regressors that directly affects D but not Y (variables excluded from the model for Y, i.e., instruments for D). We consider a bivariate triangular system for (Y, D):

$$\begin{aligned} Y &= \mathbf{1}[X'\beta + \delta_1 D - \varepsilon \geq 0], \\ D &= \mathbf{1}[X'\alpha + Z'\gamma - \nu \geq 0], \end{aligned} \tag{2.1}$$

where $\alpha \equiv (\alpha_0, \alpha_1, \ldots, \alpha_k)'$, $\beta \equiv (\beta_0, \beta_1, \ldots, \beta_k)'$, and $\gamma \equiv (\gamma_1, \gamma_2, \ldots, \gamma_l)'$. As an example of this model, Y might be an employment status or voting decision, D an indicator for having a bachelor degree, and Z college tuition. As another example, Y could be an indicator for patient death, D a medical treatment, and Z some randomization scheme. In these examples, X represents other individual characteristics.

We will maintain the following assumptions.

Assumption 1. $(X, Z) \perp (\varepsilon, \nu)$, where "$\perp$" denotes statistical independence.

Assumption 2. F_ε and F_ν are known marginal distributions of ε and ν, respectively, that are strictly increasing, are absolutely continuous with respect to Lebesgue measure, and such that $E[\varepsilon] = E[\nu] = 0$ and $Var(\varepsilon) = Var(\nu) = 1$.

Assumption 3. $(\varepsilon, \nu)' \sim F_{\varepsilon\nu}(\varepsilon, \nu) = C(F_\varepsilon(\varepsilon), F_\nu(\nu); \rho)$ where $C(\cdot, \cdot; \rho)$ is a copula known up to scalar parameter $\rho \in \Omega$ such that $C : (0, 1)^2 \to (0, 1)$ is twice differentiable in its arguments and ρ.

Assumption 4. (X', Z') does not lie in a proper linear subspace of $\mathbb{R}^{k+l}$ a.s.[5]

Assumption 1 imposes that X and Z are exogenous. This assumption, which is commonly imposed in the literature on binary choice models, excludes heteroskedasticity of the error terms. Assumption 2 characterizes the restrictions imposed on the marginal distributions of ε and ν. The moment restrictions are merely normalizations as long as the second moments of ε and ν are finite. Under these normalizations, the intercept parameter is present in the model and the correlation coefficient is the only

3 As mentioned, the results of Meango and Mourifié (2014) and Han and Lee (2017) show that an exclusion restriction is also necessary for identification when the common exogenous covariates are binary.

4 This contrasts with the identification result in a model related to ours, i.e., the sample selection model by Heckman (1979), where identification can be achieved solely by the functional form of the joint normal errors as long as there are common exogenous covariates. Excluded instruments only become necessary for identification in that model once the normality assumption is relaxed, which is not the case in our model.

5 A proper linear subspace of $\mathbb{R}^{k+l}$ is a linear subspace with a dimension strictly less than $k + l$. The assumption is that, if M is a proper linear subspace of $\mathbb{R}^{k+l}$, then $\Pr[(X', Z') \in M] < 1$.