Journal of Econometrics 177 (2013) 14–33

Contents lists available at ScienceDirect

Journal of Econometrics

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

Binary choice models with discrete regressors: Identification and misspecification

Tatiana Komarova *

London School of Economics and Political Science, United Kingdom

ARTICLE INFO

Article history:
Received 17 February 2009
Received in revised form
12 April 2013
Accepted 16 May 2013
Available online 10 June 2013

JEL classification:
C2
C10
C14
C25

Keywords:
Binary response models
Discrete regressors
Partial identification
Misspecification
Linear programming
Support vector machines

ABSTRACT

This paper explores the inferential question in semiparametric binary response models when the continuous support condition is not satisfied and all regressors have discrete support. I focus mainly on the models under the conditional median restriction, as in Manski (1985). I find sharp bounds on the components of the parameter of interest and outline several applications. The formulas for bounds obtained using a recursive procedure help analyze cases where one regressor's support becomes increasingly dense. Furthermore, I investigate asymptotic properties of estimators of the identification set. I describe a relation between the maximum score estimation and support vector machines and propose several approaches to address the problem of empty identification sets when the model is misspecified.

1. Introduction

The econometrics literature on inference in semiparametric binary response models has used support conditions on observable regressors to guarantee point identification of the vector parameter of interest. These support conditions always require continuity of one (or more) regressors. In practice though, it is not uncommon to have data sets where all regressors have discrete support, such as age, years of education, number of children and gender. In these cases, the parameter of interest is not point identified, that is, a large set of parameters may be consistent with the model. Therefore, it is important to develop methods of drawing accurate inferences without continuous support conditions on the data.

This paper examines the question of identification in semiparametric binary response models in the absence of continuity. Consider

$$Y = 1(X\beta + U \geq 0), \qquad (BR)$$

where Y is an observable binary outcome, U is an unobservable, real-valued, scalar random variable, β is a k-dimensional parameter, and X is an observable random variable with discrete support. I impose a weak median condition on the error term, as in Manski (1985):

$$M(U|X = x) = 0 \quad \text{for any } x \text{ in support of } X. \qquad (MED)$$

In this framework, I provide a methodologically new approach to the analysis of binary response models. The paper makes the following contributions.

I note that the parameter's identification region is described by a system of a finite number of linear inequalities and therefore represents a convex polyhedron. To construct this system, it is enough to know whether conditional probabilities $P(Y = 1|X = x)$ are greater or less than 0.5. As was shown by Manski and Thompson (1986), under the median condition the sign of index $x\beta$ is the same as the sign of $P(Y = 1|X = x) - 0.5$. Moreover, Manski (1988) used this fact to establish a general non-identification result for the case of discrete regressors. I exploit a recursive procedure that allows us to easily find sharp bounds on the components of the parameter of interest β. Although this procedure was outlined, for example, in Kuhn (1956) and Solodovnikov (1977), it has not been used in the context of identification.

* Tel.: +44 0 20 7852 3707; fax: +44 0 20 7955 6592.
E-mail address: t.komarova@lse.ac.uk.

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2013.05.005

I derive formulas for bounds on the components of the parameter, which prove useful in analyzing cases when the support of one regressor can become increasingly dense, that is, when we approach point identification conditions in Manski (1988). Furthermore, I show that the recursive procedure can be used not only to find sharp bounds but also to determine other characteristics of the identification region. Moreover, it can be employed in the extrapolation problem when we want to learn about $P(Y = 1|X = x_0)$ for a point x_0 that is off the support. In addition, because identification regions in ordered response and in single-index models with a monotone link function are described by systems of linear inequalities, the recursive procedure can be applied to them too.

Another contribution of the paper is to link binary response models to support vector machines (SVMs) in statistical learning theory. When the support of X is discrete and the median condition *(MED)* holds, binary response models classify points in the support into two groups and every parameter value from the identification set defines a hyperplane that separates these groups. SVMs, in their turn, is a learning technique that focuses on finding a special hyperplane that efficiently separates two classes of given training data. The major difference is that binary response models aim to find all separating hyperplanes, whereas SVMs seek only one hyperplane.

Because models might carry some degree of specification error, the recursive procedure may cease working in some situations. Therefore, it is important to develop techniques that address the consequences of model misspecification. The third contribution of this paper is to offer several methods for dealing with the issue, all of which are based on the optimization of certain objective functions. One approach is the maximum score estimation method presented in Manski (1975, 1985). Another allows us to measure the degree of misspecification by finding the minimal number of classification errors. Each method features a crucial property: the set of solutions coincides with the identification set when the model is well specified. The third approach is a modification of a soft margin hyperplane approach in SVMs and it lets us determine the extent of misspecification by determining the minimal size of a general classification error. For a well specified model, this approach gives the closure of the identification set.

Another contribution of this paper is to explore the estimation of the identification region and the sharp interval bounds. Although this paper focuses on identification, it is of interest to analyze cases where conditional probabilities $P(Y = 1|X = x)$ are not known, but their estimates $\widehat{P}(Y = 1|X = x)$ are available. In this situation, we can find estimates of identification sets and sharp interval bounds from a system of linear inequalities that uses $\widehat{P}(Y = 1|X = x)$ instead of $P(Y = 1|X = x)$. I show that when the model is well specified, such set estimators of identification sets (sharp interval bounds) converge to the true identification set (true sharp interval bounds) arbitrarily fast in terms of Hausdorff distances. I find that the sets of maximum score estimates possess the same property. I also construct confidence regions for the identification set and the sharp interval bounds and show that because of the discrete nature of the problem, they are usually conservative.

The paper presents the results of a Monte Carlo experiment with a well-specified model. The error term satisfies the median condition but is not independent of the regressors. I show that the estimates of the sharp interval bounds obtained from the system of inequalities that uses estimated conditional probabilities coincide with the identification intervals for the components of the parameter. The same is true for the sets of maximum score estimates for individual components of the parameter. For coefficients corresponding to non-constant regressors, I find the set of maximum rank correlation estimates, which turn out to lie inside the identification intervals but form much smaller sets. I also present normalized probit and logit estimates. Though these estimates are located inside the identification intervals, they are far from the value of the parameter, which was used to generate the model.

The last contribution of this paper is an empirical application which is based on data regarding the labor force participation of married women. The decision of women to participate in the labor force is treated as a dependent binary variable and regressed on education, age, labor market experience and number of children. I use different estimation techniques and compare their results. Given that misspecification or sampling error leaves the system of inequalities constructed from the estimates of conditional probabilities without solutions, I use methods suggested for dealing with the misspecification problem. I also find normalized probit and logit estimates, ordinary least squares and least absolute deviation estimates, and compare them to other estimates.

This paper is related to two strands of the literature. The first one embodies a considerable amount of work on partially identified models in econometrics. Studies on partial identification were largely initiated and advanced by Manski (see, for example, Manski, 1990, 1995, 2003), Manski and Tamer (2002) and carried further by other researchers.

The second strand analyzes models with discrete regressors. This topic is relatively underdeveloped in econometric theory, in spite of its importance for empirical work. An example of a paper that touches upon this subject is Honore and Tamer (2006). The authors describe how to characterize the identification set for dynamic random effects discrete choice models when points in the support have discrete distributions. For single-index models $E(Y|X = x) = \phi_\theta(x\theta)$ with discrete explanatory variables and no assumption on the link function ϕ_θ except for measurability, Bierens and Hartog (1988) show that there is an infinite number of observationally equivalent parameters. In particular, the identification set of the k-dimensional parameter $\theta = (\theta_1, \ldots, \theta_k)$ normalized as $\theta_1 = 1$ will be whole space $\Re^{k-1}$, with the exception of a finite number of hyperplanes (or a countable number of hyperplanes if the regressors have a discrete distribution with an infinite number of values). In binary response models with discrete regressors, Manski (1988) provides a general non-identification result and Horowitz (1998) demonstrates that the parameter can be identified only in very special cases. Magnac and Maurin (2005) also address identification issues in binary response models with discrete regressors. Their framework, however, is different from the framework in this paper. They consider a case where there is a special covariate among the regressors and assume that the model satisfies two conditions related to this covariate — partial independence and large support conditions.

The rest of the paper is organized as follows. Section 2 explains the problem and defines the identification set. Section 3 contains the mathematical apparatus and describes the recursive procedure. It also outlines applications of the recursive procedure, in particular to single-index and ordered-response models. Section 4 analyzes the case in which the discrete support of regressors grows increasingly dense. Section 5 draws an analogy between identification in bi to SVMs Section 5.2 considers misspecification issues and suggests techniques for dealing with them.

Section 6 considers the estimation of the identification set from a sample and statistical inference. Section 7 contains the results of estimations in a Monte Carlo experiment and the empirical application based on MROZ data. Section 8 concludes and outlines ideas for future research. The proofs of theorems and propositions are collected in the Appendix.

2. Partial identification

I begin by reviewing the main point identification results in the literature as well as the support conditions that guarantee point identification.