Identification of multiple equation probit models with endogenous dummy regressors

Joachim Wilde*

Department of Economics, Martin-Luther-University Halle-Wittenberg, 06099 Halle, Germany

Received 29 September 1999; accepted 25 April 2000

Abstract

This article deals with the question whether exclusion restrictions on the exogenous regressors are necessary to identify multiple equation probit models with endogenous dummy regressors. The contradictory opinions in the literature are discussed, and a simple criterion of avoiding identification problems is formulated. © 2000 Elsevier Science S.A. All rights reserved.

Keywords: Identification; Probit model; Multiple equation model; Endogenous dummy regressor

JEL classification: C35

1. Introduction

Maddala (1983, p. 122) considers a two equation probit model in which the disturbances are correlated and the binary dependent variable of the first equation is an endogenous regressor in the second equation. He states that the parameters of the second equation are not identified if there are no exclusion restrictions on the exogenous variables. Up to now this assertion has been used for empirical work (see e.g. Holly et al., 1998, p. 518). Against that Heckman (1978, p. 957) argues in a more general context that only the full rank of the (regressor) data matrix is needed to identify the parameters.

The following paper begins with the two equation case. It is demonstrated that Maddala’s argument is only valid for his simple example whereas the parameters of more realistic models are in general identified (given the full rank of the regressor matrix). Then the argumentation is generalized for the multiple equation case, and a simple criterion of avoiding the problem mentioned by Maddala is developed.

*Fax: +49-345-5527-191.
E-mail address: wilde@wiwi.uni-halle.de (J. Wilde).
2. Two equation case

Maddala (1983, p. 122) regards the following model:

\[
\begin{align*}
y_{1i}^* &= \beta_1' x_{1i} + u_{1i} \\
y_{2i}^* &= \delta_{21} y_{1i} + \beta_2' x_{2i} + u_{2i},
\end{align*}
\]

(1)

\[
y_{1i} = \begin{cases} 1 & \text{if } y_{1i}^* > 0 \\ 0 & \text{otherwise} \end{cases}, \quad y_{2i} = \begin{cases} 1 & \text{if } y_{2i}^* > 0 \\ 0 & \text{otherwise} \end{cases}
\]

(2)

\(y_{1i}^*\) and \(y_{2i}^*\) are latent variables for which only the dichotomous variables \(y_{1i}\) and \(y_{2i}\) can be observed. \(x_{1i}\) and \(x_{2i}\) are vectors of (not necessarily distinct) exogenous variables, \((u_{1i}, u_{2i})'\) is a vector of bivariate normally distributed disturbances with the usual restrictions \(\text{Var}(u_{1i}) = \text{Var}(u_{2i}) = 1\), and \(i\) is the individual index.

(1) is a special case of a linear simultaneous equation system. The recursive structure follows from the condition of logical consistency (see, e.g., Maddala, 1983, p. 118). The crucial point for the identification problem in linear simultaneous equations is that a linear combination of equations possibly contains exactly the same variables as an original equation. But for the second equation of model (1) this problem is not given. Multiplying the equations with \(l_1\) and \(l_2\) and adding up leads to

\[
\lambda_1 y_{1i}^* + \lambda_2 y_{2i}^* = \lambda_1 \delta_{21} y_{1i} + \lambda_1 \beta_1' x_{1i} + \lambda_2 \beta_2' x_{2i} + \lambda_1 u_{1i} + \lambda_2 u_{2i},
\]

\[
\Leftrightarrow y_{2i}^* = \delta_{21} y_{1i} - \frac{\lambda_1}{\lambda_2} y_{1i}^* + \frac{\lambda_1}{\lambda_2} \beta_1' x_{1i} + \beta_2' x_{2i} + \left(\frac{\lambda_1}{\lambda_2} u_{1i} + u_{2i}\right)
\]

(3)

Even if \(x_{2i}\) contains all the variables in \(x_{1i}\), (3) differs structurally from the second equation of (1) by the term \((\lambda_1/\lambda_2) y_{1i}^*\). Thus the classical identification problem does not exist. There may be only a problem because of too small variation in the data. Precisely this aspect is pointed out by the simple example of Maddala who considers the special case when \(x_{1i}\) and \(x_{2i}\) are constants. In this case there exist four unknown parameters: \(\delta_{21}\), two intercepts and the correlation parameter \(\rho\). But the likelihood function of the model includes just four different probabilities. One of these probabilities is determined through the sum restriction. As a consequence it is not possible to estimate four parameters from three remaining probabilities. In other words: The parameters are not identified. But the only reason for this result is the lack of variation in the data. In comparison to a bivariate probit model \((\delta_{21} = 0)\) without exogenous regressors which is exactly identified, the specification of an endogenous dummy regressor enlarges the number of parameters without enlarging the variation in the data.

The parameters of the model are identified, if there exists a varying exogenous regressor. Consider for instance the simple case of one exogenous regressor \(x_i\) with only two different values, say ‘1’ and ‘2’. The number of unknown parameters increases from four to six (one new slope parameter in each equation). But now there are six independent probabilities: \(P_{111}, P_{101}, P_{011}, P_{112}, P_{102}, P_{012}\), where \(P_{kni}\) denotes the (conditional) probability \(P(y_{1i} = k, y_{2i} = m|x_i = q)\). In other words: Three independent probabilities are available for each value of \(x_i\). Consequently there is sufficient variation in the data to identify the parameters. An analogous argument gives the same result for more than one varying exogenous regressor. Thus Maddala’s statement is just valid for the simple intercept model.
However for the standard case with varying exogenous regressors the full rank of the regressor matrix is sufficient for the identification of the parameters.

3. Multiple equation case

The condition of logical consistency forces a recursive structure of the model in the case of multiple equations too (Schmidt, 1981, p. 428). Note the system in the following way:

\[
\begin{align*}
    y_{1i}^* &= \beta_1' x_{1i} + u_{1i} \\
    y_{2i}^* &= \delta_{i1} y_{1i} + \beta_2' x_{2i} + u_{2i} \\
    y_{3i}^* &= \delta_{i1} y_{1i} + \delta_{i2} y_{2i} + \beta_3' x_{3i} + u_{3i} \\
    \vdots \\
    y_{Gi}^* &= \delta_{Gi1} y_{1i} + \delta_{Gi2} y_{2i} + \ldots + \delta_{Gi,G-1} y_{G-1,i} + \beta_{Gi}' x_{Gi} + u_{Gi} \\
    y_{gi} &= \begin{cases} 
        1 & \text{if } y_{gi}^* > 0 \\
        0 & \text{otherwise}
    \end{cases}, \quad g = 1, \ldots, G,
\end{align*}
\]

\[
(u_{1i}, \ldots, u_{Gi})' \text{ } G\text{-variate normally distributed with } \text{Var}(u_{gi}) = 1, \quad g = 1, \ldots, G.
\]

In analogy to the case of two equations, the classical identification problem does not exist in (4).

Assume now \( x_{gi} = 1 \) for all \( g \). In the first equation one unknown parameter is included, in the second equation two parameters are included, and so on. Adding up of these parameters and of the correlation parameters leads to

\[
\text{Number of unknown parameters} = \sum_{n=1}^{G} n + \sum_{n=1}^{G} (n - 1) = \frac{G(G + 1)}{2} + \frac{(G - 1)G}{2} = G^2. \tag{6}
\]

The likelihood function contains \( 2^G - 1 \) independent probabilities. Since \( (2^G - 1) < G^2 \) for \( 1 < G < 5 \), three and four equation models without exogenous regressors are not identified too. But as in Section 2 the existence of varying exogenous regressors avoids this identification problem: One exogenous variable with two values leads to 14 independent probabilities in the three equation case and to 30 independent probabilities in the four equation case. It follows

**Lemma.** The existence of one varying exogenous regressor in each equation is sufficient to avoid small variation identification problems in multiple equation probit models with endogenous dummy regressors.

Remark that the above result was worked out under the implicit assumption of a multivariate normal distribution. If this assumption is dropped, the probabilities may depend on the unknown parameters in a different way and additional assumptions may be necessary for identifying the parameters.
4. Conclusion

In contrast to linear simultaneous equations with only continuous endogenous variables in recursive multiple equation probit models with endogenous dummy regressors no exclusion restrictions for the exogenous variables are needed if there is sufficient variation in the data. The last condition is ensured by the assumption that each equation contains at least one varying exogenous regressor, an assumption which is rather weak in economic applications.

References