ARTICLE IN PRESS

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

Contents lists available at ScienceDirect

Journal of Econometrics

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

Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables

Jeffrey M. Wooldridge *

Department of Economics, Michigan State University, 486 W. Circle Drive, 110 Marshall-Adams Hall, East Lansing, MI 48824-1038, United States

ARTICLE INFO

Article history:
Available online xxxx

JEL classification:
C13
C21
C25

Keywords:
Quasi-maximum likelihood
Control function
Linear exponential family
Average structural function
Variable addition test

ABSTRACT

I propose a quasi-maximum likelihood framework for estimating nonlinear models with continuous or discrete endogenous explanatory variables. Joint and two-step estimation procedures are considered. The joint procedure is a quasi-limited information maximum likelihood procedure, as one or both of the log likelihoods may be misspecified. The two-step control function approach is computationally simple and leads to straightforward tests of endogeneity. In the case of discrete endogenous explanatory variables, I argue that the control function approach can be applied with generalized residuals to obtain average partial effects. I show how the results apply to nonlinear models for fractional and nonnegative responses.

1. Introduction

It is well known that endogeneity of explanatory variables is generally difficult to handle in nonlinear models, although several special cases have been worked out. Unlike with a linear model with constant coefficients, where two stage least squares (2SLS) can always be applied regardless of the nature of the endogenous explanatory variables (EEVs), with nonlinear models the probabilistic nature of the EEVs – whether they are continuous, discrete, or some combination – plays a critical role. Methods where fitted values obtained in a first stage are plugged in for the EEVs in a second stage are generally inconsistent for both the structural parameters and other quantities of interest, such as average partial (or marginal) effects.

A common approach to estimate nonlinear models with EEVs is to apply maximum likelihood (conditional on the exogenous variables). In principle, joint MLE is available when a distribution (conditional on exogenous variables) for the EEVs is fully specified and a distribution of the response variable conditional on the EEVs (and exogenous variables) is specified or derived from a set of equations with unobserved errors. The MLE approach has been widely applied, especially for binary responses, but it has some limitations. For one, it can be computationally difficult with multiple EEVs or many sources of heterogeneity. Perhaps more importantly, it maintains specification of a full set of conditional distributions. In some cases we may wish to specify only certain features of a conditional distribution, such as a conditional mean.

Since the influential work of White (1982), econometricians have known that the estimators obtained from maximum likelihood estimation of misspecified models generally converge to parameters that can be given an interpretation in terms of best fit to the true density, and it is possible to perform inference on those parameters. Further, there are special cases where the quasi-MLE (QMLE) actually identifies population parameters that index some correctly specified feature of the distribution. Gourieroux et al. (1984) (GMT) consider the important case of conditional means and conditional variances. An important result of GMT is that a QMLE in the linear exponential family identifies a correctly specified conditional mean with essentially arbitrary distributional misspecification.

The first contribution of this paper is to show that a class of quasi-MLE methods can be used to consistently estimate parameters in nonlinear models with endogenous explanatory variables. I rely on the results of GMT along with a common partitioning of quasi-log-likelihood functions for models with EEVs. Two practically important examples are quasi-MLEs obtained for a fractional response (a variable that takes values in the unit interval) with either a continuous or a binary EEV. Conveniently, the log likelihood function for a binary response can be applied to fractional response variables under a conditional mean assumption without further restricting the conditional distribution. Naturally, because

* Tel.: +1 517 353 5972; fax: +1 517 432 1068.
E-mail address: wooldri1@msu.edu.

http://dx.doi.org/10.1016/j.jeconom.2014.04.020
0304-4076/

the method is quasi-MLE, robust inference needs to be used because the information matrix equality is generally violated. These days, such inference is routinely available.

The second contribution of this paper is to use the joint quasi-MLE framework to obtain simple variable addition tests (VATs) that can detect endogeneity of suspected explanatory variables. In particular, I show how the score principle leads to a general class of test statistics obtained by adding generalized residuals, obtained in a first-stage QMLE problem, to a second-stage quasi-MLE problem and using a standard Wald test. The VATs are easily made robust to distributional failures other than correct specification of the conditional mean. Further, because the VATs are asymptotically equivalent to score tests under correct distributional specification, the VATs are asymptotically efficient in some leading cases and can be expected to work well generally.

The VAT approach to test endogeneity is closely related to a second popular approach to estimate nonlinear models with EEVs: the two-step control function (CF) approach. In the most common CF approaches, residuals from a first-stage estimation involving reduced forms for the EEVs are inserted into a second-stage estimation problem. Popular examples are Rivers and Vuong (1988) for probit models and Smith and Blundell (1986) for Tobit models. Wooldridge (2010) uses the control function (CF) approach in a variety of settings, including nonlinear models with cross section data or panel data. In an important work, Blundell and Powell (2003, 2004) (BP) have shown that the approach has broad applicability in semiparametric and even nonparametric settings. BP show that quantities of interest – partial effects of the so-called average structural function – are identified very generally, without distributional or functional form restrictions.

The main drawback of most CF approaches for nonlinear models – even in BP's general setting – is that the nature of the EEVs is restricted. It must be assumed that the reduced forms of the EEVs have additive errors that are independent of the variables exogenous in the structural equation. The assumption of additive, independent errors rules out discrete EEVs. Thus, while the BP approach allows for general response functions, its scope is restricted because it does not allow general EEVs. The third contribution of this paper – and one that is somewhat controversial – is to recommend two-step CF approaches in a quasi-MLE framework for general kinds of EEVs. In effect, I suggest that a flexible two-step control function approach used to obtain VATs might yield good estimates of average partial effects fairly generally.

The CF approach for nonlinear models with discrete (as well as continuous) EEVs has been previously proposed by Terza et al. (2008) (TBR), which they call "two-stage residual inclusion". My derivation of the CF approach is somewhat different from TBR's. In particular, before specifying any restrictions on functional form, I use a conditional independence assumption and I focus on average partial effects rather than parameter estimates. By doing so I am able to argue for more flexible functional forms for the conditional mean while interpreting the approach as providing an approximate solution to the endogeneity problem. Plus, I consider a more general quasi-MLE estimation framework, and I discuss why the CF approach is more convincing for continuous EEVs than for discrete EEVs.

The rest of the paper is organized as follows. In Section 2 I use a standard linear model as motivation for QMLE by showing the robustness of the Gaussian limited information maximum likelihood (LIML) estimator. The arguments in the linear case can be extended to nonlinear cases, and Section 3 lays out the general approach. Section 4 shows how the approach can be applied to fractional response variables and nonnegative responses with an exponential mean function, including count responses.

Simple variable addition tests for testing the null that the suspected EEVs are exogenous are derived in Section 5. These tests are easily obtained using standard software, and they motivate the general control function approach in Section 6 for handling endogeneity of continuous and discrete EEVs. Section 6.5 contains some concluding remarks.

2. Motivation for quasi-MLE: a linear model

Consider a population linear model for a response variable, y_1, with a single endogenous explanatory variable (EEV), y_2:

$$y_1 = \alpha_{o1} y_2 + \mathbf{z}_1 \boldsymbol{\delta}_{o1} + u_1, \tag{1}$$

where $\mathbf{z}_1$ is a $1 \times L_1$ strict subvector of a vector $\mathbf{z}$. Assume the vector $\mathbf{z}$ is exogenous in the sense that

$$E(\mathbf{z}' u_1) = \mathbf{0}. \tag{2}$$

In practice, $\mathbf{z}_1$ would include a constant, and so we assume that u_1 has a zero mean. I use the convention of putting "o" on the parameters in (1) because it is helpful to distinguish the population values from generic values in the parameter space.

The reduced form of y_2 is a linear projection in the population:

$$y_2 = \mathbf{z}\boldsymbol{\delta}_{o2} + v_2 \tag{3}$$

$$E(\mathbf{z}' v_2) = \mathbf{0} \tag{4}$$

where $\boldsymbol{\delta}_{o2}$ is $L \times 1$. Notice that nothing about the linear projection defined by (3) and (4) restricts the nature of y_2; it could be a continuous variable but also a discrete variable, including a binary variable. Also, (1) can be viewed as a linear approximation to a underlying linear model, where (2) effectively defines α_{o1} and $\boldsymbol{\delta}_{o1}$.

Provided $E(\mathbf{z}'\mathbf{z})$ is nonsingular and $\boldsymbol{\delta}_{o22} \neq \mathbf{0}$, where $\boldsymbol{\delta}_{o2} = (\boldsymbol{\delta}_{o21}', \boldsymbol{\delta}_{o22}')'$, two stage least squares (2SLS) estimation under random sampling is consistent; see, for example, Wooldridge (2010, Chapter 5). An alternative approach, and one that is convenient for testing the null that y_2 is exogenous, is a control function approach. Write the linear projection of u_1 on v_2, in error form, as

$$u_1 = \gamma_{o1} v_2 + e_1, \tag{5}$$

where $\gamma_{o1} = E(v_2 u_1)/E(v_2^2)$ is the population regression coefficient. By construction, $E(v_2 e_1) = 0$ and $E(\mathbf{z}' e_1) = \mathbf{0}$.

If we plug (5) into (1) we can write

$$y_1 = \alpha_{o1} y_2 + \mathbf{z}_1 \boldsymbol{\delta}_1 + \gamma_{o1} v_2 + e_1 \tag{6}$$

$$E(\mathbf{z}' e_1) = \mathbf{0}, \qquad E(v_2 e_1) = 0, \qquad E(y_2 e_1) = 0. \tag{7}$$

Adding the reduced form error, v_2, to the structural equation "controls" for the endogeneity of y_2. If we could observe data on v_2, we could simply add it as a regressor. Instead, given a random sample of size N, we can estimate $\boldsymbol{\delta}_{o2}$ in a first stage by OLS and obtain the residuals, $\hat{v}_{i2}, i = 1, \ldots, N$. In a second stage we run the regression

$$y_{i1} \text{ on } y_{i2}, \mathbf{z}_{i1}, \text{ and } \hat{v}_{i2}, \quad i = 1, \ldots, N. \tag{8}$$

The OLS estimators from (8) are control function (CF) estimators. It is well known – for example, Hausman (1978) – that the estimates $\hat{\alpha}_1$ and $\hat{\boldsymbol{\delta}}_1$ are *identical* to the 2SLS estimates. See also Wooldridge (2010, Chapter 6).

Rather than use a two-step method, an alternative is to obtain the LIML estimator assuming that (u_1, v_2) is independent of $\mathbf{z}$ and bivariate normal, which implies that (e_1, v_2) is bivariate normal and independent of $\mathbf{z}$. For variance parameters η_1^2 and τ_2^2, the log likelihood for random draw i (conditional on $\mathbf{z}_i$), multiplied by two, is

$$-\log(\eta_1^2) - [y_{i1} - \alpha_1 y_{i2} - \mathbf{z}_{i1}\boldsymbol{\delta}_1 - \gamma_1(y_{i2} - \mathbf{z}_i\boldsymbol{\delta}_2)]^2/\eta_1^2 - \log(\tau_2^2) - (y_{i2} - \mathbf{z}_i\boldsymbol{\delta}_2)^2/\tau_2^2, \tag{9}$$

Please cite this article in press as: Wooldridge, J.M., Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. Journal of Econometrics (2014), http://dx.doi.org/10.1016/j.jeconom.2014.04.020