Economics Letters 152 (2017) 41–45

Contents lists available at ScienceDirect

Economics Letters

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

Endogeneity bias modeling using observables

Antonio F. Galvao [a], Gabriel Montes-Rojas [b], Suyong Song [c,*]

[a] Department of Economics, University of Iowa, W284 Pappajohn Business Building, 21 E. Market Street, Iowa City, IA 52242, United States
[b] CONICET-Universidad de San Andrés, Vito Dumas 284, Victoria, B1644, Provincia de Buenos Aires, Argentina
[c] Department of Economics, University of Iowa, W360 Pappajohn Business Building, 21 E. Market Street, Iowa City, IA 52242, United States

HIGHLIGHTS

- This paper proposes an alternative solution to the endogeneity problem.
- Endogeneity bias is modeled as a function of additional observables.
- Identification of the parameters of interest is provided.
- We propose an estimator and show its consistency and asymptotic normality.

ARTICLE INFO

Article history:
Received 26 February 2016
Received in revised form 10 September 2016
Accepted 15 December 2016
Available online 29 December 2016

JEL classification:
C10
C26

Keywords:
Endogeneity
Instrumental variables
Proxy variables

ABSTRACT

This paper proposes an alternative solution to the endogeneity problem by explicitly modeling the joint interaction of the endogenous variables and the unobserved causes of the dependent variable as a function of additional observables. We derive identification of the parameters, develop an estimator, and establish its consistency and asymptotic normality.

1. Introduction

The problem of endogeneity occupies a substantial amount of research in theoretical and applied econometrics. The most popular solutions are instrumental variables (IV) (see e.g. Hausman, 1983; Angrist and Krueger, 2001 for surveys) and proxy variables approach (see e.g. Olley and Pakes, 1996; Levinsohn and Petrin, 2003). These solutions rely on exogenous information derived from an additional exclusion restriction. In applications, the type of restriction chosen determines the nature of the model to be used, i.e. the instrument or the *proxy* variable. However, in many empirical applications, there is frequently disagreement and concern about the exclusion restrictions imposed, and instruments and proxies selections. The potential IV are often argued to be invalid since they are still correlated with the error term (see, e.g., Bound et al. (1995) and Hahn and Hausman (2002)) while the conditions for identification using proxy variables are many times implausible.

Recently there has been an expanding literature on analyzing endogeneity when IV and proxy variables models fail. This literature explores alternative moment conditions and exclusion restrictions. For instance, Altonji et al. (2005a, b, 2008) develop a strategy to extract information from observables about the endogeneity bias. They construct an index of observables, which can be used to identify the endogenous variable parameter, in combination with prior knowledge about the sign of the bias and a condition on the relationship between included (observable) and excluded (non-observable) variables. Chalak and White (2011) define a new class of extended IV, and introduce notions of conditioning and conditional extended IV which allow use of non-traditional instruments, as they may be endogenous. Chalak (2012) achieves identification of parameters by employing restrictions on the magnitude and sign of confounding instead of using traditional IV. Nevo and Rosen (2012) provide bounds for the parameters when the standard exogeneity assumption on IV fails, by assuming the correlation

* Corresponding author.
E-mail addresses: antonio-galvao@uiowa.edu (A.F. Galvao), gmontesrojas@udesa.edu.ar (G. Montes-Rojas), suyong-song@uiowa.edu (S. Song).

http://dx.doi.org/10.1016/j.econlet.2016.12.021
0165-1765/

between the instruments and the error term has the same sign as the correlation between the endogenous regressor and the error term and that the instruments are less correlated with the error term than is the endogenous regressor. Montes-Rojas and Galvao (2014) exploit information on the structure of endogeneity and use prior information in a Bayesian framework to infer about the potential heterogeneity in parameter estimators.

This paper proposes an alternative solution to the endogeneity problem by explicitly modeling the joint interaction of the endogenous variables and the unobserved causes of the dependent variable as a function of additional observables, defined as *simultaneous variables*. Identification uses the endogeneity structure of the model to build an alternative moment condition which is based on the non-zero conditional expectation implied by the endogeneity. That is, rather than imposing a sign on the endogeneity effect or exploring the bounds derived from its potential magnitude, we work with an alternative moment restriction. The intuition on the main identification condition of the new procedure is that, by using the proposed condition, the econometrician is able to model the endogeneity bias using the additional observable variables. Our framework allows for situations in which there are no valid standard IV or proxy variables available, but there exist additional variables that happen to be related to both the endogenous variable and the unobserved causes of the dependent variable. We develop a simple estimator based on the identification, and establish its consistency and asymptotic normality.

Many potential empirical applications might benefit from the proposed approach, especially those where the potential IV might still be related to the unobservables, or the proposed proxy variable does not satisfy all the requirements. Consider the errors-in-variables setting to motivate its empirical relevance. Many empirical applications rely on lagged mismeasured variables as IV to solve the implied endogeneity (see e.g. Biorn, 2000). The validity of the IV would fail if the measurement error is persistent because the instruments (i.e. lagged mismeasured variables) would still be correlated with the error term. More reliable estimates could be obtained by modeling the joint interaction of the mismeasured variable and the error term as a function of lagged mismeasured variables (see e.g. Galvao et al., 2016).

The paper is organized as follows. Section 2 presents the econometric model and establishes identification. Section 3 develops a consistent estimator and establishes its asymptotic properties.

2. The model

Consider the following structural model

$$y_i = \mathbf{x}_{1i}\boldsymbol{\beta}_1 + \mathbf{x}_{2i}\boldsymbol{\beta}_2 + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}_1$ is a p_1-vector, $\boldsymbol{\beta}_2$ is a p_2-vector, and ϵ_i is a scalar innovation term. Define $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top]^\top$. We assume that $\mathbf{x}_{2i}$ is endogenous, and correlated with the innovation term ϵ_i in (1), such that $E[\mathbf{x}_{2i}^\top \epsilon_i] \neq 0$. In addition, $\mathbf{x}_{1i}$ is exogenous with $E[\mathbf{x}_{1i}^\top \epsilon_i] = \mathbf{0}$. The endogeneity in $\mathbf{x}_2$ produces endogeneity bias. To solve the endogeneity problem we will model the interaction of the endogenous variable and the error term, $\mathbf{x}_{2i}^\top \epsilon_{i\top}$, and establish identification of $\boldsymbol{\beta}$ under mild conditions. For simplicity, throughout we consider the case where $p_2 = 1$, i.e., there is only one endogenous variable, x_{2i}. Extension to the multivariate case is straightforward.

The following equation formalizes modeling endogeneity,

$$E(x_2\epsilon \mid \mathbf{z}, \mathbf{x}) = \mathbf{z}\boldsymbol{\phi}. \tag{2}$$

Eq. (2) considers a linear model only for simplicity, but it could be extended to a nonparametric model (e.g., method of sieve). It explicitly models the endogeneity of x_2 using additional variables $\mathbf{z}$, defined as *simultaneous variables*. In this case, by modeling endogeneity we mean to model the term $x_2\epsilon$. When $\boldsymbol{\phi} \neq \mathbf{0}$, we can interpret the exogenous variable $\mathbf{z}$ as a noisy measure of the common cause(s) of x_2 and ϵ, which is related to the *joint* interaction of the endogenous variable and the unobservables. Our identification strategy requires observable variables, $\mathbf{z}$.

The proposed identification is related to the control function approach. When the correlation between x_2 and ϵ is modeled, Eq. (2) can be rewritten as $x_2 E(\epsilon \mid \mathbf{z}, \mathbf{x}) = \mathbf{z}\boldsymbol{\phi}$, hence, we have $E(\epsilon \mid \mathbf{z}, \mathbf{x}) = \frac{\mathbf{z}}{x_2}\boldsymbol{\phi}$. Therefore, the conditional expected value of the unobserved error term is a function of the "normalized" variables , i.e., $\frac{\mathbf{z}}{x_2}$. The emphasis is however on the nature of $\mathbf{z}$, which provides information about the joint interaction of the endogenous variable and the error term.

We are interested in identifying and estimating the parameters $\boldsymbol{\beta}$ in Eq. (1). In practice, $\boldsymbol{\phi}$ is unknown, and it is important to note that this parameter cannot be directly estimated from Eq. (2) because ϵ is unobservable. Define $\boldsymbol{\theta} \equiv [\boldsymbol{\beta}_1^\top, \boldsymbol{\alpha}^\top]^\top$ with $\boldsymbol{\alpha} \equiv [\beta_2^\top, \boldsymbol{\phi}^\top]^\top$. To ease the notation, define $\widetilde{y}$ and $\widetilde{x}_2$ after netting out the exogenous regressor $\mathbf{x}_1$ and multiplying the resulting objects by x_2. Thus, $\widetilde{y} = x_2(y - \mathbf{x}_1 E(\mathbf{x}_1^\top \mathbf{x}_1)^{-1} E(\mathbf{x}_1^\top y))$ and $\widetilde{\mathbf{x}} = [\widetilde{x}_2, \mathbf{z}]$, with $\widetilde{x}_2 = x_2(x_2 - \mathbf{x}_1 E(\mathbf{x}_1^\top \mathbf{x}_1)^{-1} E(\mathbf{x}_1^\top x_2))$. Let $\widetilde{\mathbf{z}}$ be a set of variables induced by conditioning variables $[\mathbf{z}, \mathbf{x}]$. Note that in this case we are obtaining the residual projection on $\mathbf{x}_1$. Consider the following assumptions.

Assumption 1.

(*i*) $E(\mathbf{x}_1^\top \epsilon) = \mathbf{0}$;
(*ii*) $E(x_2\epsilon \mid \mathbf{z}, \mathbf{x}) = \mathbf{z}\boldsymbol{\phi}$.

Assumption 2. $E(\mathbf{x}_1^\top \mathbf{x}_1)$ and $E(\widetilde{\mathbf{z}}^\top \widetilde{\mathbf{x}})$ are non-singular.

Assumptions 1 and 2 allow identification of the parameters of interest. Assumption 1(*i*) simply states that $\mathbf{x}_1$ are exogenous regressors. Assumption 1(*ii*) is the main identification condition. It is new in the literature and deserves further discussion. Assumption 1(*ii*) explicitly models the interaction between the endogenous variable and the unobserved causes of the dependent variable using a parametric model specification. It states that $\mathbf{z}$ is able to capture the information on the endogeneity term. The intuition behind this assumption is that once one controls for $\mathbf{z}$, $\mathbf{x}$ is not related to the interaction term $x_2\epsilon$. In other words, the endogeneity bias implied by the non-zero conditional expectation of the interaction term can be specified as a function of $\mathbf{z}$.

It is important to notice the restrictions this assumption imposes relative to the IV approach in the literature. For simplicity we consider a model with only one (endogenous) covariate $y = x\beta + \epsilon$. In our case, the additional equation can be rewritten as $x\epsilon = z\phi + u$ where u is the orthogonal projection of $x\epsilon$ on z. Our required moment conditions are two: $E[zu] = 0$ and $E[x^2 u] = 0$. The IV model requires dependence between the endogenous regressor and the instrumental variables, which are restricted to be uncorrelated with the error term. This could be written as an additional equation $x = z\phi + u$ (where now u is the orthogonal projection of x on z) with also two moment conditions $E[zu] = 0$ and $E[z\epsilon] = 0$. Our method is able to allow the additional variable(s) z to still be correlated with the error term, ϵ, and also the endogenous variable to be correlated with u, the residual (unexplained) component in the additional equation. As a result, the difference between our proposed model and traditional IV approach rests on different model specifications; researchers fail to identify parameters if an incorrect method is employed to control for the endogeneity in each case. In our case we model endogeneity, the correlation of x and ϵ, i.e. $x\epsilon$.

We now return to the general structural equation (1) and general identification. For the sake of clarity, we focus on exactly