Economics Letters 124 (2014) 100–103

Contents lists available at ScienceDirect

Economics Letters

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

A simple estimator for partial linear regression with endogenous nonparametric variables[☆]

Michael S. Delgado[a,*], Christopher F. Parmeter[b,1]

[a] Department of Agricultural Economics, Purdue University, West Lafayette, IN 47907, United States

[b] Department of Economics, University of Miami, Coral Gables, FL 33124, United States

HIGHLIGHTS

- We focus on semiparametric regression with endogeneity.
- Our estimator is simple to implement.
- Our estimator performs well in finite samples.

ARTICLE INFO

Article history:
Received 22 February 2014
Received in revised form
31 March 2014
Accepted 28 April 2014
Available online 9 May 2014

Keywords:
Semiparametric
Partial linear
Endogeneity
Instrumental variables
Monte Carlo

ABSTRACT

We propose a simple kernel estimator for semiparametric partial linear models with endogeneity in the nonparametric function. Compared to the existing backfitting estimator, our estimator is notationally simpler and relatively easier to implement. We also discuss data-driven bandwidth selection to implement this estimator in practice. Monte Carlo exercises show that the finite sample performance of these two estimators is similar.

1. Introduction

Recently, researchers have considered variants of the nonparametric triangular system of equations setup rigorously studied by Newey et al. (1999). In particular, Su and Ullah (2008) propose a kernel regression estimator for the fully nonparametric specification in Newey et al. (1999) while Martins-Filho and Yao (2012) consider a kernel regression estimator for a semiparametric partial linear variant of the same specification.[2] The result has been the development of practical tools that applied researchers can deploy in a straightforward fashion.

In this paper, we focus on estimation of $m(\cdot)$ and β in the partial linear specification

$$Y_i = m(X_{1i}) + X_{2i}\beta + \varepsilon_i \tag{1}$$

in which Y_i is a scalar outcome variable, X_{1i} and X_{2i} are d_1- and d_2-dimensioned vectors of conditioning variables, $m(\cdot) : \mathbb{R}^{d_1} \rightarrow \mathbb{R}$ is a smooth function of X_{1i}, β is a d_2-dimensioned vector of parameters, ε_i is a scalar disturbance term, and the index $i = 1, 2, \ldots, n$ denotes the sample. Further assume that for any variable in X_1,

$$X_{1i} = g(Z_i) + U_i \tag{2}$$

for a p-dimensioned vector of variables Z_i, some smooth function $g(\cdot) : \mathbb{R}^p \rightarrow \mathbb{R}$, and scalar disturbance U_i. We follow Newey

☆ This research was supported in part by computational resources provided by Information Technology at Purdue—Rosen Center for Advanced Computing, Purdue University, West Lafayette, Indiana. All R code used in this paper is available upon request.

* Corresponding author. Tel.: +1 765 494 4211.
E-mail addresses: delgado2@purdue.edu (M.S. Delgado), cparmeter@bus.miami.edu (C.F. Parmeter).

1 Tel.: +1 305 284 4397.

2 We focus on these papers in particular because they explicitly assume identical conditional moment restrictions to Newey et al. (1999). For similar models, albeit with differences in assumed conditional moment restrictions that are not necessarily more or less general than those considered here, see, for example, Ai and Chen (2003) and Otsu (2011).

http://dx.doi.org/10.1016/j.econlet.2014.04.032

et al. (1999) and Martins-Filho and Yao (2012) and impose the conditional moment restrictions $E[\varepsilon_i|X_{1i}] \neq 0$, $E[U_i|Z_i] = 0$, and $E[\varepsilon_i|Z_i, U_i] = E[\varepsilon_i|U_i]$, so that X_{1i} is endogenous and Z_i serves as a proper instrumental variable.

Given the partial linear restriction on (1), our specification is similar to the model considered by Martins-Filho and Yao (2012) and is the well-known specification made famous by Robinson (1988). Our discussion here will focus on the existence of endogeneity in the nonparametric function, and not in the linear component X_{2i} as well. In the event that X_{1i} is correlated with ε_i, standard estimation approaches (e.g., Robinson, 1988) cannot yield consistent estimates without modification. See, for example, Li and Stengos (1996) for an estimator of a partial linear model with endogeneity in the parametric part.

Martins-Filho and Yao (2012) propose using a control function approach (Newey et al., 1999; Su and Ullah, 2008) to eliminate the endogeneity in (1), followed by a backfitting estimator to estimate $m(\cdot)$ and β. They choose not to deploy a marginal integration estimator (e.g., Su and Ullah, 2008) in order to obtain more efficient estimates (Kim et al., 1999). Monte Carlo exercises demonstrate that their estimation approach performs well in finite samples. Further, the marginal integration estimator of, for example, Su and Ullah (2008) is unnecessary here if we were to assume a parametric form for $g(Z_i)$ in (2), as demonstrated in Blundell and Duncan (1998).

While it is clear that the Martins-Filho and Yao (2012) backfitting estimator is theoretically superior to a marginal integration approach, we contend that their approach may be somewhat cumbersome for applied researchers to implement given the relative notational complexity and iterative steps required to estimate $m(\cdot)$ and β. Further, while the backfitting approach requires less computational time, recent advances in parallel computing render computational time less of an obstacle for applied work (Delgado and Parmeter, 2013). The purpose of this research is to study the relative finite sample performance of both the marginal integration approach (described below) and the Martins-Filho and Yao (2012) approach. In our view, the marginal integration approach is relatively simpler both in terms of notational burden and implementation, and may be more accessible for applied researchers. Hence, we seek an assessment of the relative finite sample performance of both estimators. We further provide a discussion of data-driven bandwidth selection for our estimator, as data-driven methods are usually considered necessary in applied settings (Li and Racine, 2007).

We highlight that Model (1) is a popular choice for applied econometricians who seek to incorporate flexibility into their regression specification while avoiding dimensionality issues common in fully nonparametric specifications. Currently, the basic partial linear specification (without endogeneity) has been applied in the context of economic growth (e.g., Liu and Stengos, 1999), environmental economics (e.g., Millimet et al., 2003), and consumer demand (e.g., Blundell et al., 1998).

2. Estimation strategy

2.1. Estimator

Here we describe the marginal integration approach for estimating (1). We refer the reader to Martins-Filho and Yao (2012) for a derivation of the backfitting approach. Following Newey et al. (1999), Su and Ullah (2008) and Martins-Filho and Yao (2012), under the conditional moment restrictions assumed above and using the Law of Iterated Expectations:

$$E[Y_i|X_{1i}, X_{2i}, Z_i, U_i] = m(X_{1i}) + X_{2i}\beta + E[\varepsilon_i|U_i]. \tag{3}$$

This implies a model of the form

$$Y_i = m(X_{1i}) + X_{2i}\beta + E[\varepsilon_i|U_i] + \upsilon_i \tag{4}$$

in which υ_i is defined to be purely random error: $\upsilon_i = Y_i - E[Y_i|X_{1i}, X_{2i}, Z_i, U_i]$. The insight from past research (e.g., Newey et al., 1999) is to define $E[\varepsilon_i|U_i] = h(U_i)$ for some function $h(\cdot) : \mathbb{R}^{d_1} \to \mathbb{R}$, and replace the unknown U_i with an estimate from the reduced form regression of X_{1i} on Z_i : $\widehat{U}_i = X_{1i} - \widehat{g}(Z_i)$. The model then becomes

$$Y_i = m(X_{1i}) + X_{2i}\beta + h(\widehat{U}_i) + \upsilon_i. \tag{5}$$

Our proposed estimator is as follows. First, reformulate the model as

$$Y_i = m_0(X_{1i}, \widehat{U}_i) + X_{2i}\beta + \upsilon_i \tag{6}$$

and then apply the conditional mean transformation of Robinson (1988) to recover estimates of $m_0(\cdot)$ and β. It is possible to apply the Robinson (1988) estimator to (6) but not (1) given the presence of $\widehat{U}_i$ that controls for the endogeneity of X_{1i}. That is, using a nonparametric estimator, obtain estimates of $E[Y_i|X_{1i}, \widehat{U}_i]$ and $E[X_{2i}|X_{1i}, \widehat{U}_i]$ and construct the model

$$Y_i^\star = X_{2i}^\star\beta + \upsilon_i^\star \tag{7}$$

in which $Y_i^\star = Y_i - E[Y_i|X_{1i}, \widehat{U}_i]$, $X_{2i}^\star = X_{2i} - E[X_{2i}|X_{1i}, \widehat{U}_i]$, and $\upsilon_i^\star = \upsilon_i - E[\upsilon_i|X_{1i}, \widehat{U}_i]$. Ordinary least squares can be used to regress $Y_i^\star$ on $X_{2i}^\star$ to recover an estimate of β, which can then be used to construct

$$\widetilde{Y}_i = m_0(X_{1i}, \widehat{U}_i) + \widetilde{\upsilon}_i \tag{8}$$

in which $\widetilde{Y}_i = Y_i - X_{2i}\widehat{\beta}$. A nonparametric estimator can be used to obtain an estimate of $m_0(\cdot)$, from which we recover an estimate of $m(\cdot)$ via marginal integration

$$\widehat{m}(X_{1i}) = n^{-1}\sum_{j=1}^{n}\widehat{m}_0(X_{1i}, \widehat{U}_j). \tag{9}$$

This final step, in particular, lends easily to recent advances in parallel computing accessible to most applied researchers (Delgado and Parmeter, 2013). Further, Henderson and Parmeter (2014) show that one need not integrate over the entire sample to obtain reliable estimates of $m(\cdot)$.[3] Throughout, we advocate using a local-linear least-squares estimator to estimate unknown conditional means.

2.2. Bandwidth selection

We advocate the use of data-driven methods to recover the bandwidths to estimate all nonparametric functionals in this procedure, including both $g(Z_i)$ and $m(X_{1i})$. An obvious choice in this endeavor is least-squares cross-validation (Li and Racine, 2007). Our cross-validation criterion function is:

$$\min_{h_2}\sum_{i=1}^{n}\left[y_i - \widehat{m}_{-i}(X_{1i})\right]^2, \tag{10}$$

where $\widehat{m}_{-i}(X_{1i})$ is the leave-one-out estimator of $m(X_{1i})$, and h_2 is the vector of bandwidths used to construct $m_{-i}(X_{1i})$. Note that the bandwidths h_1 used to construct $\widehat{U}_i$ via $\hat{g}(Z_i)$ in (2) are also calculated in this procedure, however, as noted in Su and Ullah (2008), these bandwidths do not effect the asymptotic performance of $\widehat{m}(X_1)$.

[3] In particular, Henderson and Parmeter (2014) show that using roughly 25% of the overall sample produces estimates with almost identical bias and variance as that using the full sample.