Journal of Econometrics 157 (2010) 165–178

Contents lists available at ScienceDirect

Journal of Econometrics

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

A semiparametric cointegrating regression: Investigating the effects of age distributions on consumption and saving[☆]

Joon Y. Park[a,b,*], Kwanho Shin[c], Yoon-Jae Whang[d]

[a] *Department of Economics, Indiana University, United States*
[b] *School of Economics, Sungkyunkwan University, Republic of Korea*
[c] *Department of Economics, Korea University, Republic of Korea*
[d] *Department of Economics, Seoul National University, Republic of Korea*

ARTICLE INFO

Article history:
Available online 31 October 2009

JEL classification:
C22

Keywords:
Cointegration
Semiparametric regression
Series estimation
Age distributions
Consumption and savings rate

ABSTRACT

We consider a semiparametric cointegrating regression model, for which the disequilibrium error is further explained nonparametrically by a functional of distributions changing over time. The paper develops the statistical theories of the model. We propose an efficient econometric estimator and obtain its asymptotic distribution. A specification test for the model is also investigated. The model and methodology are applied to analyze how an aging population in the US influences the consumption level and the savings rate. We find that the impact of age distribution on the consumption level and the savings rate is consistent with the life-cycle hypothesis.

1. Introduction

Nonlinear and nonparametric models have drawn much attention over the past decade, and it now seems generally agreed that many important economic relationships are intrinsically complex and cannot be modeled effectively using simple linear parametric models. See, for example, Pagan and Ullah (1999) and Granger and Teräsvirta (1993). Many empirical economic relationships, which were formerly represented by linear and parametric models, are indeed being modeled and estimated within nonlinear and nonparametric frameworks. Nonlinear and nonparametric approaches certainly provide more flexibility and accommodate a broader class of models. The generality, however, comes at a cost. It is well known that estimated nonlinear and nonparametric models often have relatively poor finite sample performance and/or slower rates of convergence asymptotically. The cost can be prohibitively high in nonstationary time series settings, as is well demonstrated in Park (2005), and this forces us to look for reasonable compromises. The class of partially nonlinear and nonparametric models certainly offers one such compromise.

We consider in this paper a semiparametric model that is partially nonlinear and nonparametric. The linear part of the model specifies a cointegrating relationship among a set of integrated variables in a parametric regression form, whereas the nonlinear part describes the effect of a functional regressor which we model nonparametrically by introducing a response function. In the paper, we consider a series estimation of the model and establish its statistical theories. In particular, we develop an efficient econometric estimator of the model and obtain its asymptotic distribution. A specification test, which we may use to check the adequacy of the model, is also introduced and analyzed. The efficient estimator is asymptotically Gaussian, and the proposed specification test has a limit chi-square distribution. Our statistical theories in the paper are therefore all Gaussian. The assumptions used in the paper are mild and allow for a wide class of integrated processes and functional regressors that may appear in practical applications.

The model and methodology developed in the paper can be used to analyze various time series macroeconomic models from a new perspective. In particular, our approach allows for modeling economic relationships over time that are also affected by the cross-sectional distributions in each period. Typically, individual specific variables such as age are averaged out when we investigate the relationships among a set of aggregate variables. Consequently,

☆ We are grateful to a Guest Editor and two anonymous referees for many useful comments. We also like to thank Yoosoon Chang, Oliver Linton and Akiko Tamura for helpful discussions and comments, and Qi Li and Kwan Koo Yun for their patience, encouragement and support. This research is supported financially by a Hosei University Research Grant, which we gratefully acknowledge.

* Corresponding author at: Department of Economics, Indiana University, United States.

E-mail address: jpark@tamu.edu (J.Y. Park).

0304-4076/$ – see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.jeconom.2009.10.032

we overlook the interaction between the levels of aggregate macroeconomic variables and their distributions since we expect that the distribution of age itself also matters in determining the levels of the aggregate macroeconomic variables. For example, the life cycle hypothesis suggests that the savings rate should vary along the stages of the life-cycle. Then the aggregate savings rate should depend on the age distribution, in addition to other typical aggregate macroeconomic variables such as the level of income or income growth.

When we apply our methodology to analyze the quarterly consumption level and savings rate from 1959:1 to 2002:3, we find that the impact of age distribution on the consumption level and the savings rate takes a U-shape and an inverted U-shape respectively, which is consistent with the life-cycle hypothesis. In other words, we find that consumption is lower for middle-agers and higher for both sides of young-agers and old-agers, while the savings rate responds in a mirrored fashion to the age distribution, being highest for middle-agers. This is in contrast with the previous studies based on parametric approaches, which repeatedly produced a U-shaped impact curve for the savings rate. We point out the reasons why our nonparametric approach outperforms the conventional parametric approaches. Of course, our nonparametric approach also has some well known costs: a slower rate of convergence for the estimates of model parameters and the dependency of inferential results on the choice of basis functions and truncation parameters.

The rest of the paper is organized as follows. Section 2 introduces the model and assumptions. The details of the model are given with the required assumptions, and the series estimation method is introduced to estimate the model. The basic statistical theories are developed in Section 3. Various approximation results are provided to facilitate the asymptotic analysis of our model. The limit distributions of the model estimators are also given, and the asymptotics of the long-run error variance estimator are developed. Section 4 provides the efficient estimation method based on the CCR transformation, and the specification test using the variable addition approach that suits our model conveniently well. Section 5 reports the empirical results for the application of the model; these analyze the effect of age distribution on the consumption level and savings rate. Section 6 concludes the paper, and the mathematical proofs are in the Appendix.

A word on notation. As usual, we use $|\cdot|$ to denote the modulus. If applied to vectors or matrices, the notation denotes the maximum of the moduli of their components. For a vector $x = (x_i)$, $\|\cdot\|$ signifies the standard Eucleadian norm, i.e., $\|x\|^2 = \sum x_i^2$. On the other hand, the notation is used to denote the operator norm for a matrix $A = (a_{ij})$. We therefore have $\|A\| = \sup \|Ax\|/\|x\|$. It is well known that $\|A\|^2$ is dominated by the maximum eigenvalue of $A'A$, and consequently, bounded in particular by $\|A\|^2 \leq \mathrm{tr}(A'A) = \sum a_{ij}^2$. We also use the same notation, $\|\cdot\|$, to signify the supremum norm for continuous functions defined on a compact interval. This should cause no confusion. For functions that are vector-valued, the notation denotes the maximum of the supremum norms of the component functions. Standard notations such as o_p and O_p for stochastic orders, and $\to_p$ and $\to_d$ for convergences of random sequences, are used without any reference. Moreover, equality in distribution is denoted by $=_d$, and $\mathbb{R}$ denotes the set of real numbers. Finally, we denote by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ respectively the smallest and largest eigenvalues of a matrix A.

2. The model and assumptions

We consider the regression model given by

$$y_t = \nu_t + x_t'\beta + u_t, \tag{1}$$

where

$$\nu_t = \int_{-\infty}^{\infty} f_t(s)g(s)\mathrm{d}s \tag{2}$$

for $t = 1, \ldots, n$. The model (1) is partially linear, consisting of both linear and nonlinear parts. The regressor (x_t) has a linear relationship with the regressand (y_t), which is specified parametrically. On the other hand, in the nonlinear part (ν_t), the regressor (f_t) is given as a set of functional observations, to which the regressand (y_t) responds nonlinearly as given in (2). The function g may be interpreted as a *response function*, which measures the effect of (f_t) on (y_t) in a nonparametric fashion. Throughout the paper, we assume that the functional regressor (f_t) is deterministic.[1] As usual, (u_t) denotes the regression error.

In what follows, we assume that (x_t) is a vector integrated process and (u_t) a stationary process, so that our model (1) represents a semiparametric cointegrating regression. Let (x_t) be m-dimensional. Further, we let $v_t = \Delta x_t$,

$$w_t = (u_t, v_t')', \tag{3}$$

and assume the following.

Assumption 1. Let

$$w_t = \sum_{i=0}^{\infty} \Phi_i \varepsilon_{t-i}$$

and we assume that $\Phi(1)$ is nonsingular and $\sum i|\Phi_i| < \infty$, and that (ε_t) are iid with $\Psi = \mathbb{E}\varepsilon_t\varepsilon_t' > 0$ and $\mathbb{E}|\varepsilon_t|^p < \infty$ for some $p > 4$.

The process (w_t) is therefore assumed to be a general linear process. We only require the standard summability condition for (Φ_i) and the moment condition for (ε_t), which are routinely imposed in the time series literature. The iid assumption on the innovation sequence (ε_t) is restrictive, since it does not allow for more general linear processes driven by ARCH-type innovations having conditional heteroskedasticity. The assumption does not appear to be crucial, and is imposed here to ease the proofs of our subsequent theoretical results.

Let

$$a = (p-2)/2p,$$

where p is the maximal order of the existing moment for (ε_t) introduced in Assumption 1. Under Assumption 1, we have by the result in Park and Hahn (1999):

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{[nr]} w_t = \Phi(1)\frac{1}{\sqrt{n}}\sum_{t=1}^{[nr]}\varepsilon_t + O_p(n^{-a})$$

uniformly in $r \in [0, 1]$. Therefore, due to the strong approximation obtained by, e.g., Einmahl (1989), we have the following.

Lemma 1. *Let Assumptions 1 and 2 hold. Then there exists a stochastic process W_n defined on $[0, 1]$ such that*

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{[nr]} w_t =_d W_n(r) \tag{4}$$

and

$$\sup_{0\leq r\leq 1} |W_n(r) - W(r)| = O_p(n^{-a}),$$

where W is a vector Brownian motion with variance $\Omega = \Phi(1)\Psi\Phi(1)'$.

[1] We may of course let (f_t) be random and given exogenously.