Journal of Econometrics 187 (2015) 238–255

Contents lists available at ScienceDirect

Journal of Econometrics

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

Estimation in generalised varying-coefficient models with unspecified link functions

Wenyang Zhang[a], Degui Li[a,*], Yingcun Xia[b,c]

[a] Department of Mathematics, University of York, United Kingdom

[b] Department of Statistics and Applied Probability, National University of Singapore, Singapore

[c] School of Mathematical Sciences, University of Electronic Science and Technology of China, China

ARTICLE INFO

Article history:
Received 15 April 2013
Received in revised form
6 October 2014
Accepted 16 February 2015
Available online 28 February 2015

JEL classification:
C13
C14

Keywords:
Generalised varying-coefficient models
Identifiability
Iterative estimation procedure
Kernel smoothing
Weighted least squares

ABSTRACT

In this paper, we study the generalised varying-coefficient models, where the link function is unspecified and the response variable can be either continuous or discrete. As the link function is unspecified, the models under investigation become unidentifiable. In this paper, we derive an identification condition for the generalised varying-coefficient models, which is much weaker and more reasonable than that given by Kuruwita et al. (2011) whose model can be seen as a special case of our modelling framework. Under the identification condition, we introduce a nonparametric iterative procedure to estimate the functional coefficient with its direction and norm as well as the unspecified link function, and then establish the asymptotic properties of the resulting nonparametric estimators. Furthermore, a weighted least squares based algorithm is provided to implement the iterative estimation procedure. The simulation studies and empirical application show that our estimation methodology works quite well in both small and median sample cases.

1. Introduction

Varying-coefficient models are important to explore the nonlinear dynamic pattern in data analysis, and have experienced rapid development in both theoretical and applied aspects; see, for example, Fan and Zhang (1999), Li et al. (2002), Zhang et al. (2002), Xia et al. (2004), Cai (2007), Sun et al. (2007), Cheng et al. (2009), Wang et al. (2009), Zhang et al. (2009), Li and Racine (2010) and Li and Zhang (2011). Extensions of the varying-coefficient models include varying-coefficient single-index models (Lu et al., 2006; Kuruwita et al., 2011) and generalised varying-coefficient models (GVCMs; Cai et al., 2000; Zhang and Peng, 2010; Zhang, 2011). As in the generalised linear models (GLMs), the existing literature usually assumes that the link function is specified in the GVCMs, and the specification is often somewhat arbitrary. In some practical applications, the commonly-used link functions might be questionable or even misleading. Thus, a data-driven approach to select the link function is imperative. For the GLMs, a special case of GVCMs, there exists some literature on how to choose the link function by a data-driven approach, see, for example, Pregibon (1980), McCullagh and Nelder (1989), Mallick and Gelfand (1994), Weisberg and Welsh (1994), and Carroll et al. (1995). However, to the best of our knowledge, there is no data-driven approach to select the link function in the GVCMs. Kuruwita et al. (2011) did use a data-driven approach to select the link function in their model. However, their model is in fact a varying-coefficient single-index model rather than the GVCMs, which will be discussed in detail in Section 2.

Once the link function is unspecified, the GVCMs would become very complicated and unidentifiable. It is not desirable to borrow the identification condition for single-index models (Horowitz and Härdle, 1996; Carroll et al., 1997) and simply restrict the first component of the functional coefficient to be positive and the norm of the vector of the functional coefficient to be one to make the models identifiable, see, for example, Kuruwita et al. (2011). Such an identification condition makes the models unrealistic and limits their application, which will be explained later. This paper aims to relax this restrictive condition and derive an identification condition which is weaker and more reasonable, and then introduce an iterative procedure to estimate the nonparametric components in the GVCMs. In order to achieve notational economy, we use GVCMUL to denote the GVCMs with unknown link functions. The main contributions of the paper can be summarised as follows.

[*] Correspondence to: Department of Mathematics, University of York, York, YO10 5DD, United Kingdom.
E-mail address: degui.li@york.ac.uk (D. Li).

http://dx.doi.org/10.1016/j.jeconom.2015.02.022
0304-4076/

(i) For the developed framework of GVCMUL, we allow that the response variable can be either continuous or discrete, whereas Kuruwita et al. (2011) only considered the case of continuous response. Furthermore, making use of the modelling structure, we derive a novel identification condition for the GVCMUL, which does not require point-wise identifiability for the function (Kuruwita et al., 2011). Our identification condition presented is not only a matter of mathematical consideration, but also fundamentally enlarges the scope of the application of the models and provides a new way to estimate the unknown functions in the models.

(ii) The estimation of the unknown functions in GVCMUL is challenging as the so-called "curse of dimensionality" issue (Fan and Yao, 2003) becomes very acute under such circumstance. In this paper, we propose an estimation procedure based on the developed identification condition of GVCMUL and an iterative procedure, which could avoid the dimensionality problem, to estimate the functional coefficient, the direction and norm of the functional coefficient, and the link function. Furthermore, we introduce a weighted least squares based algorithm to implement the proposed estimation procedure. In contrast, it is not very easy to implement the method in Kuruwita et al. (2011) even under their identification condition, because the minimisation in Step 1 of their method could be too difficult to implement and its accuracy is questionable.

(iii) Under some regularity conditions, we establish the asymptotic properties of the nonparametric estimators obtained by the iterative estimation procedure. Furthermore, we conduct some simulation studies to examine the finite sample performance of our estimation procedure, and the simulation results show that our estimation procedure works quite well. In particular, through studying the same simulated example as that in Kuruwita et al. (2011), we show that our method performs much better than the method in Kuruwita et al. (2011). We finally apply the proposed GVCMUL and the estimation procedure to analyse some environmental data from Hong Kong. The empirical results highlight the importance of estimating the link function from the data.

The rest of the paper is organised as follows. The description of the GVCMUL and the identification condition are given in Section 2. The nonparametric estimation procedure is introduced in Section 3, and the asymptotic theory is provided in Section 4. The performance of the proposed estimation procedure is assessed by some simulation studies and an empirical application in Section 5, where the comparison between our estimation procedure and the method in Kuruwita et al. (2011) is also presented. Section 6 concludes the paper. The assumptions and proofs of the asymptotic results are given in an Appendix.

2. Model and identifiability

We next introduce the modelling structure and derive the model identification condition. Let y be a response variable of interest, which can be either continuous or discrete. When y is discrete, we define the density function of y as its probability mass function. Let U be a scalar covariate, $\mathbf{X}$ a p-dimensional vector of covariates, and $m(U, \mathbf{X}) = \mathbb{E}(y \mid U, \mathbf{X})$ the conditional expectation of y given U and $\mathbf{X}$. Assume that the log conditional density function of y given U and $\mathbf{X}$ is

$$C_1(\boldsymbol{\phi}_1)\log f\big(m(U, \mathbf{X}), y\big) + C_2(y, \boldsymbol{\phi}_2) \quad \text{with } m(U, \mathbf{X}) = g\big(\mathbf{X}^{\mathrm{T}}\mathbf{a}(U)\big), \tag{2.1}$$

where $f(\cdot, \cdot)$, $C_1(\cdot)$ and $C_2(\cdot, \cdot)$ are known, $C_1(\boldsymbol{\phi}_1) > 0$, $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are unknown nuisance parameters, neither the link function $g(\cdot)$ nor the functional coefficient $\mathbf{a}(\cdot) = \big[a_1(\cdot), \ldots, a_p(\cdot)\big]^{\mathrm{T}}$ is known. The modelling framework (2.1) is motivated by the commonly-used assumption in both econometrics and statistics that the conditional density of y given U and $\mathbf{X}$ belongs to the exponential family, which is shown in the following example.

Example 2.1. Let the conditional density of y given U and $\mathbf{X}$ belong to the exponential family defined by

$$\exp\left\{\frac{y\cdot\eta(U, \mathbf{X}) - \mathcal{B}_1\big(\eta(U, \mathbf{X})\big)}{\mathcal{A}(\boldsymbol{\phi})} + \mathcal{B}_2(y, \boldsymbol{\phi})\right\}, \tag{2.2}$$

where the functions $\mathcal{A}(\cdot)$, $\mathcal{B}_1(\cdot)$ and $\mathcal{B}_2(\cdot, \cdot)$ are known, $\boldsymbol{\phi}$ is the dispersion parameter and $\eta(\cdot, \cdot)$ is similar to the natural parameter in the context of the parametric GLMs which carries the information from U and $\mathbf{X}$. If we further assume that $\eta(U, \mathbf{X})$ can be expressed as $\eta(U, \mathbf{X}) = \mathbf{X}^{\mathrm{T}}\mathbf{a}(U)$, the conditional expectation $m(U, \mathbf{X})$ can be linked to $\mathcal{B}_1\big(\eta(U, \mathbf{X})\big)$ via

$$m(U, \mathbf{X}) = \mathcal{B}_1'\big(\eta(U, \mathbf{X})\big) = \mathcal{B}_1'\big(\mathbf{X}^{\mathrm{T}}\mathbf{a}(U)\big), \tag{2.3}$$

where $\mathcal{B}_1'(\cdot)$ is the derivative of $\mathcal{B}_1(\cdot)$. The above equation indicates that $\eta(U, \mathbf{X})$ can be written as a function $m(U, \mathbf{X})$. Noting that $\mathcal{A}(\cdot)$, $\mathcal{B}_1(\cdot)$ and $\mathcal{B}_2(\cdot, \cdot)$ are assumed to be known, we can thus show that the log conditional density function in the exponential family can be written in the form given in (2.1). The link function $g(\cdot)$ (which is unknown in our modelling structure and estimated by a data-driven method), would reduce to $\mathcal{B}_1'(\cdot)$ in the exponential distribution family. We next consider two commonly-used distributions in the exponential family to provide more explicit explanation.

- When the conditional distribution of y given U and $\mathbf{X}$ is a normal distribution with the density function defined by

  $$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\big[y - m(U, \mathbf{X})\big]^2}{2\sigma^2}\right\},$$

  we can derive (2.2) by letting $\boldsymbol{\phi} = \sigma^2$, $\mathcal{A}(z) = z$, $\mathcal{B}_1(z) = z^2/2$, $\mathcal{B}_2(z, \boldsymbol{\phi}) = \mathcal{B}_2(z, \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{z^2}{2\sigma^2}$ and $\eta(U, \mathbf{X}) = m(U, \mathbf{X})$. The parameter σ^2 can be treated as a nuisance parameter, if the main interest is to estimate the conditional mean function of y given U and $\mathbf{X}$. Furthermore, if $\eta(U, \mathbf{X}) = \mathbf{X}^{\mathrm{T}}\mathbf{a}(U)$, (2.3) reduces to $m(U, \mathbf{X}) = \mathbf{X}^{\mathrm{T}}\mathbf{a}(U)$, which is the form of the varying-coefficient model. Hence, we can show that this case falls into the modelling framework (2.1) by treating σ^2 as a nuisance parameter.
- When the conditional distribution of y given U and $\mathbf{X}$ is a Poisson distribution with the probability mass function defined by

  $$\mathbb{P}(y = k \mid U, \mathbf{X}) = \frac{m^k(U, \mathbf{X})}{k!}\exp\big\{-m(U, \mathbf{X})\big\}, \quad k = 0, 1, 2, \ldots,$$

  we can derive (2.2) by letting $\mathcal{A}(\boldsymbol{\phi}) \equiv 1$, $\mathcal{B}_1(z) = e^z$, $\mathcal{B}_2(k; \boldsymbol{\phi}) = -\log(k!)$ and $\eta(U, \mathbf{X}) = \log m(U, \mathbf{X})$. Furthermore, if $\eta(U, \mathbf{X}) = \mathbf{X}^{\mathrm{T}}\mathbf{a}(U)$, (2.3) reduces to

  $$m(U, \mathbf{X}) = \exp\big\{\eta(U, \mathbf{X})\big\} = \exp\big\{\mathbf{X}^{\mathrm{T}}\mathbf{a}(U)\big\}.$$

  Thus, we can show that the modelling framework (2.1) is satisfied by taking $g(z) = e^z$.

Like in the GLMs, the main interest of this paper lies in the conditional mean of the response variable, and $C_1(\boldsymbol{\phi}_1)$ and $C_2(y, \boldsymbol{\phi}_2)$ have little to do with the mean component. Hence, without loss of generality, we assume in this paper that the log conditional density function of y given U and $\mathbf{X}$ is

$$\log f\,(m(U, \mathbf{X}), y) \quad \text{with } m(U, \mathbf{X}) = g\left(\mathbf{X}^{\mathrm{T}}\mathbf{a}(U)\right). \tag{2.4}$$

It is worthwhile to compare model (2.4) with the model studied in Kuruwita et al. (2011):

$$y = g\left(\mathbf{X}^{\mathrm{T}}\mathbf{a}(U)\right) + \epsilon, \qquad \mathbb{E}(\epsilon|\mathbf{X}, U) = 0, \tag{2.5}$$