Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 901–914
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Generalized partially linear varying-coefficient models ☆

Yiqiang Lu

Institute of Electronic Technology, PLA Information Engineering University, Zhengzhou, China

Received 30 December 2004; received in revised form 9 February 2007; accepted 28 February 2007
Available online 25 April 2007

Abstract

Generalized varying-coefficient models are useful extensions of generalized linear models. They arise naturally when investigating how regression coefficients change over different groups characterized by certain covariates such as age. In this paper, we extend these models to generalized partially linear varying-coefficient models, in which some coefficients are constants and the others are functions of certain covariates. Procedures for estimating the linear and non-parametric parts are developed and their associated statistical properties are studied. The methods proposed are illustrated using some simulations and real data analysis.

MSC: primary 62G05; secondary 62G08; 62G20

Keywords: Generalized linear models; Varying-coefficient models; Partially linear varying-coefficient models; Local polynomial regression; Asymptotic theory

1. Introduction

Generalized linear models (GLMs), which have extremely wide applications, provide a framework for relating response and predictor variables (McCullagh and Nelder, 1989). In many practical situations, however, GLMs are not flexible enough to capture the underlying relationship between the response variable and its associated covariates. In recent years, various attempts have been made to relax these model assumptions and to exploit possible hidden structures. Examples include additive models (Breiman and Friedman, 1985; Hastie and Tibshirani, 1990), partially linear models (Wahba, 1984; Green and Silverman, 1994), generalized partially linear single-index models (Carroll et al., 1997), low-dimensional interaction models, (Friedman, 1991; Gu and Wahba, 1993; Stone et al., 1997) and multiple-index models (Härdle and Stoker, 1989; Li, 1991), among others. Of importance are the varying-coefficient models, in which the coefficients of GLMs are replaced by smoothing non-parametric functions and hence the regression coefficients are allowed to vary as functions of other factors (Hastie and Tibshirani, 1993).

The varying-coefficient model has the form

$$g(\mu(x, u)) = \sum_{j=1}^{p} \alpha_j(u) x_j \tag{1.1}$$

☆ This research was supported by a grant from the National Natural Science Foundation of China (No. 10501053.)
E-mail address: yiqiang_lu@163.com.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2007.02.010

for some given link function $g(\cdot)$, where $x=(x_1,\ldots,x_p)^\tau$, and $\mu(x,u)$ is the mean regression function of the response variable Y given the covariates $U=u$ and $X=x$ with $X=(X_1,\ldots,X_p)^\tau$. (By regarding $X_1\equiv 1$, model (1.1) allows a varying intercept term.) Varying-coefficient models are useful extensions of classical GLMs. These extensions allow simple interpretation. The appeal of model (1.1) is that by allowing coefficients $\alpha_1,\ldots,\alpha_p$ to depend on U, the modeling bias can be reduced significantly and the "curse of dimensionality" can be avoided. The varying-coefficient model arises in many different contexts and has been successfully applied to GLMs, time series analysis, longitudinal data analysis, and environmental data analysis. See, for example, the work of Hoover et al. (1998), Cai et al. (2000a,b) and Chen and Tsay (1993), among others.

In practice, investigators often want to know whether a covariate affects the response or whether coefficients are really varying (Fan and Zhang, 2000; Cai et al., 2000b). This amounts to testing whether the whole function is zero or constant, namely, testing the null hypothesis H_0 that $\alpha_j(u)=\beta_j$, for some j. The model under H_0 is called a generalized partially linear varying-coefficient model (GPLVCM), which is defined by the following linear model:

$$g(\mu(x,z,u))=\sum_{j=1}^{p}x_j\beta_j+\sum_{j=1}^{q}z_j\alpha_j(u)=x^\tau\beta+z^\tau\alpha(u), \tag{1.2}$$

where $x=(x_1,\ldots,x_p)^\tau$, $z=(z_1,\ldots,z_q)^\tau$, $\beta=(\beta_1,\ldots,\beta_p)^\tau$, $\alpha(u)=(\alpha_1(u),\ldots,\alpha_q(u))^\tau$ and $\mu(x,z,u)$ is the mean regression function of the response variable Y given the covariates $U=u$, $X=x$, and $Z=z$, with $X=(X_1,\ldots,X_p)^\tau$ and $Z=(Z_1,\ldots,Z_q)^\tau$. This model consists of a linear part that involves constant coefficients β_j, $j=1,\ldots,p$, and a non-parametric part that involves coefficient functions $\alpha_j(u)$, $j=1,\ldots,q$. The classical generalized partially linear model (Carroll et al., 1997; Green and Silverman, 1994) can be regarded as a special case with $Z_1\equiv 1$, $Z_2=\cdots=Z_q=0$. In the absence of $x^\tau\beta$, (1.2) becomes the varying-coefficient models. On the other hand, if the constant coefficient β_j is viewed as a function, the GPLVCM (1.2) can be regarded as a special case of the generalized varying-coefficient model (1.1). β and $\alpha(\cdot)$ can be estimated by maximizing the local quasi-likelihood. Since the coefficient functions of (1.2) clearly allow different degrees of smoothness, the two-step estimation method should be used to obtain estimates of β and $\alpha(\cdot)$.

In the least-square setting, (1.2) with an identity link was introduced by Zhang et al. (2002), and further studied by Xia et al. (2004), in which a semi-local least-squares estimation procedure was proposed. Fan and Huang (2005) investigated the testing problem for the parametric component. In this paper, our aim is to estimate the unknown parameter β and the unknown coefficient function $\alpha(\cdot)$ in the full model (1.2) with a known link function $g(\cdot)$. Our work also applies to quasi-likelihood models in which only the relationship between the mean and the variance is specified. In this situation, estimation of the mean can be achieved by replacing the log-likelihood condition $\log f_{Y|X,Z,U}(y,x,z,u)$ by a quasi-likelihood function $Q(\mu(x,z,u),y)$. If the conditional variance is modeled as $\operatorname{var}(Y|X,Z,U)=\sigma^2V(\mu(x,z,u))$ for some known positive function $V(\cdot)$, the corresponding quasi-likelihood function $Q(w,y)$ satisfies:

$$\frac{\partial}{\partial w}Q(w,y)=(y-w)/V(w) \tag{1.3}$$

(McCullagh and Nelder 1989, Chapter 9). Quasi-score (1.3) possesses properties similar to those of the usual likelihood score function.

The article is organized as follows. In Section 2, the estimation procedures are proposed. Section 3 describes the distribution theory. In Section 4, the performance of the proposed methods is illustrated by some simulation studies and a real example. Section 5 presents a discussion of the non-parametric coefficient function that allows different degrees of smoothness. Finally, proofs are given in Appendix A.

2. Estimation methods

The primary interest of model (1.2) is for estimating β and $\alpha(\cdot)$. Since $\alpha(\cdot)$ is modeled non-parametrically, it is natural to consider the local quasi-likelihood. However, efficient estimation of the global parameter β requires the use of all data points, and hence should rely on the global quasi-likelihood. Thus, the following two-step estimation method is proposed.