Computational Statistics and Data Analysis 117 (2018) 45–61

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Optimal QR-based estimation in partially linear regression models with correlated errors using GCV criterion

Mahdi Roozbeh

Faculty of Mathematics, Statistics and Computer Science, Semnan University, P.O. Box 35195-363, Semnan, Iran

ARTICLE INFO

Article history:
Received 9 April 2016
Received in revised form 22 July 2017
Accepted 2 August 2017
Available online 9 August 2017

Keywords:
Generalized ridge estimation
Kernel smoothing
Multicollinearity
QR decomposition
Partially linear regression model
Shrinkage parameter

ABSTRACT

Multicollinearity among the predictor variables is a serious problem in regression analysis. There are some classes of biased estimators for solving the problem in statistical literature. In these biased classes, estimation of the shrinkage parameter plays an important role in data analyzing. Using eigenvalue analysis, efforts have been made to develop skills and methods for computing risk function of the estimators in regression models. A modified estimator based on the QR decomposition to combat the multicollinearity problem of design matrix is proposed in partially linear regression model which makes the data to be less distorted than the other methods. The necessary and sufficient condition for the superiority of the partially generalized QR-based estimator over partially generalized least-squares estimator for selecting the shrinkage parameter is obtained. Under appropriate assumptions, the asymptotic bias and variance of the proposed estimators are obtained. Also, a generalized cross validation (GCV) criterion is proposed for selecting the optimal shrinkage parameter and the bandwidth of the kernel smoother and then, an extension of the GCV theorem is established to prove the convergence of the GCV mean. Finally, the Monté-Carlo simulation studies and a real application related to electricity consumption data are conducted to support our theoretical discussion.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Partially linear regression models (PLRMs) are appropriate models when a suitable link function of the mean response is assumed to have a linear parametric relationship to some explanatory variables while its relationship to the other variables has an unknown form. Let $(y_1, x_1^\top, t_1), \ldots, (y_n, x_n^\top, t_n)$ be the observations that follow the partially Linear regression model, that is,

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + f(t_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where $\boldsymbol{x}_i^\top = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a vector of explanatory variables, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is an unknown p-dimensional vector parameter, t_i's are design points which belong to some bounded domain $D \subset \mathbb{R}$, $f(t)$ is an unknown smooth function and ϵ_i's are random errors which are assumed to be independent of $(\boldsymbol{x}_i, t_i)$.

Surveys regarding the estimation and application of the model (1.1) can be found in the monograph of Härdle et al. (2000). Speckman (1988) studied partial residual estimation of $\boldsymbol{\beta}$ and $f(\cdot)$ in (1.1), and obtained asymptotic bias and variance of the estimators. He showed that these estimators are less biased compared to the partial smoothing spline estimators. Bunea (2004) proposed a consistent covariate selection technique in an PLRM through penalized least-squares criterion. He showed that the selected estimator of the linear part is asymptotically normal. You and Chen (2007) considered the problem of

E-mail address: mahdi.roozbeh@semnan.ac.ir.

http://dx.doi.org/10.1016/j.csda.2017.08.002
0167-9473/© 2017 Elsevier B.V. All rights reserved.

estimation in model (1.1) with serially correlated errors, obtained the semiparametric generalized least-squares estimator of the parametric component and studied the asymptotic properties of it. You et al. (2007) developed statistical inference for the model (1.1) for both heteroscedastic and/or correlated errors under general assumption $Var(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, with a positive definite matrix $\boldsymbol{V}$, is supposed to hold. For bandwidth selection in the context of kernel-based estimation in model (1.1), Li et al. (2011) used cross-validation criteria for optimal bandwidth selection.

In regression analysis, researchers often encounter the problem of multicollinearity that is defined as the existence of nearly linear dependency among columns of the design matrix $\boldsymbol{X}$. The existence of multicollinearity may lead to wide confidence intervals for the individual parameters or linear combination of the parameters and may produce estimates with wrong signs.

The most popular approach to combat multicollinearity is the ridge regression estimator proposed by Hoerl and Kennard in the 1970s. Several other methods for dealing with multicollinearity are the r–k class estimator proposed by Baye and Parker (1984), the biased estimator proposed by Liu (1993) and the r–d class estimator proposed by Kaçiranlar and Sakallioğlu (2001). As a brief review on applicability of these strategies, Akdenïz and Tabakan (2009), Akdenïz Duran et al. (2011), Kibria and Saleh (2011), Roozbeh et al. (2011), Akdenïz Duran and Akdenïz (2012), Roozbeh and Arashi (2013), Amini and Roozbeh (2015), Arashi and Valizadeh (2015), Arashi et al. (2015), and Roozbeh (2015, 2016) employed ridge methodology in facing with partially Linear regression model. Nomura and Ohkubo (1985), and Sarkar (1996) considered r–k class estimator. Hubert and Wijekoon (2006), Yang et al. (2009), Yang and Xu (2011), and Arashi et al. (2014) used Liu's approach.

The main part of this paper is devoted to overcome multicollinearity using the QR decomposition and study the asymptotic properties of the partial residual QR-based estimator of $\boldsymbol{\beta}$ and $f(.)$ in model (1.1) with correlated errors. This work is organized as follows: contains some usual estimation methods used for estimating the ridge parameter in partially Linear regression models, together with a modified one. In Section 4, a class of QR estimators is studied and then, its properties are extracted and superiority condition of the new estimator in contrast to the partial generalized least-squares estimator is given. Section 5 is devoted to obtaining the asymptotic bias and variance of the proposed estimators. To select the optimal bandwidth of the kernel smoother and shrinkage parameters, the generalized cross validation criteria are proposed in Section 6. An extension of the GCV theorem of Golub et al. (1979) is established to prove the convergence of the expectation of the GCV criterion. Section 7 is devoted to the Monté-Carlo simulation studies and an application in bridge construction data. Finally, conclusions are drawn in Section 8.

2. The partial residual ridge estimation

Consider the following partially Linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f}(\boldsymbol{t}) + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ is a $n \times p$ matrix, $\boldsymbol{f}(\boldsymbol{t}) = (f(t_1), \ldots, f(t_n))^\top$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$. We assume that in general, $\boldsymbol{\epsilon}$ is a vector of disturbances, which is distributed with $\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{E}(\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, where σ^2 is an unknown parameter and $\boldsymbol{V}$ is a symmetric, positive definite known matrix.

To estimate $\boldsymbol{\beta}$ and $f(t)$ for a point $t \in D$, first consider the simplified model

$$\boldsymbol{y} = \boldsymbol{f}(t) + \boldsymbol{\epsilon}, \tag{2.2}$$

obtained from (2.1) with $\boldsymbol{\beta} = \boldsymbol{0}$. The linear smoother of $f(t)$ in (2.3) is $\hat{f}(t) = \boldsymbol{k}_{\omega_n}(t)\boldsymbol{y}$, with $\boldsymbol{k}_{\omega_n}(t) = (K_{\omega_n}(t, t_1), \ldots, K_{\omega_n}(t, t_n))$, where $K_{\omega_n}(.)$ is a kernel function of order m with bandwidth parameter ω_n. For the existence of $\hat{f}(t, \boldsymbol{\beta})$ at the optimal convergence rate $n^{-4/5}$, in partially Linear regression models with probability one, we need some conditions on kernel function. See Müller (2000) for more details.

If the kernel function $K_{\omega_n}(.)$ is of order m, according to Speckman (1988), there exist bounded functions h_1 and h_2, such that for each $t \in D$,

$$\mathrm{E}\big(\boldsymbol{k}_{\omega_n}(t)\boldsymbol{y} - f(t)\big) = \omega_n^m h_1(t) f^{(m)}(t) + o(\omega_n^m), \tag{2.3}$$

and

$$\mathrm{Cov}\big(\boldsymbol{k}_{\omega_n}(t)\boldsymbol{y}\big) = \sigma^2(n\omega_n)^{-1}h_2(t)(1 + o(1)), \tag{2.4}$$

where $f^{(m)}(t)$ is the mth derivative of $f(t)$.

To estimate the parameters of the model (2.1), we first remove the non-parametric effect, apparently. Assuming $\boldsymbol{\beta}$ to be known, a natural nonparametric estimator of $f(.)$ is $\hat{f}(t) = \boldsymbol{k}_{\omega_n}(t)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Replacing $f(t)$ by $\hat{f}(t)$ in (2.1), the model is simplified to

$$\check{\boldsymbol{y}} = \check{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.5}$$

where $\check{\boldsymbol{y}} = \big(\boldsymbol{I}_n - \boldsymbol{K}_{\omega_n}\big)\boldsymbol{y}$, $\check{\boldsymbol{X}} = \big(\boldsymbol{I}_n - \boldsymbol{K}_{\omega_n}\big)\boldsymbol{X}$ and $\boldsymbol{K}_{\omega_n}$ is the smoother matrix with i, jth component $K_{\omega_n}(t_i, t_j)$.

Download English Version:

https://daneshyari.com/en/article/4949186

Download Persian Version:

https://daneshyari.com/article/4949186

Daneshyari.com