Journal of Multivariate Analysis 147 (2016) 127–144

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Robust ridge estimator in restricted semiparametric regression models

Mahdi Roozbeh

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

ARTICLE INFO

Article history:
Received 15 June 2015
Available online 28 January 2016

AMS subject classifications:
primary 62G08
62G35
secondary 62J05
62J07

Keywords:
Breakdown point
Generalized restricted ridge estimator
Kernel smoothing
Least trimmed squares estimator
Linear restrictions
Multicollinearity
Outlier
Robust estimation
Semiparametric regression model

ABSTRACT

In this paper, ridge and non-ridge type estimators and their robust forms are defined in the semiparametric regression model when the errors are dependent and some non-stochastic linear restrictions are imposed under a multicollinearity setting. In the context of ridge regression, the estimation of shrinkage parameter plays an important role in analyzing data. Another common problem in applied statistics is the presence of outliers in the data besides multicollinearity. In this respect, we propose some robust estimators for shrinkage parameter based on least trimmed squares (LTS) method. Given a set of n observations and the integer trimming parameter $h \leq n$, the LTS estimator involves computing the hyperplane that minimizes the sum of the smallest h squared residuals. The LTS estimator is closely related to the well-known least median squares (LMS) estimator in which the objective is to minimize the median squared residual. Although LTS estimator has the advantage of being statistically more efficient than LMS estimator, the computational complexity of LTS is less understood than LMS. Here, we extract the robust estimators for linear and nonlinear parts of the model based on robust shrinkage estimators. It is shown that these estimators perform better than ordinary ridge estimator. For our proposal, via a Monté-Carlo simulation and a real data example, performance of the ridge type of robust estimators are compared with the classical ones in restricted semiparametric regression models.

1. Introduction

Semiparametric regression models (SRMs) 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 semiparametric 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 $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{ip})$ is a vector of explanatory variables, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ 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. [14]. Speckman [38] 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.

E-mail addresses: mahdi.roozbeh@profs.semnan.ac.ir, m.roozbeh.stat@gmail.com.

http://dx.doi.org/10.1016/j.jmva.2016.01.005

Bunea [10] proposed a consistent covariate selection technique in an SRM through penalized least squares criterion. He showed that the selected estimator of the linear part is asymptotically normal. You and Chen [42] considered the problem of 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. [43] developed statistical inference for the model (1.1) for both heteroscedastic and/or correlated errors under general assumption $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, 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. [22] used cross-validation criteria for optimal bandwidth selection.

Now consider a semiparametric regression model in the presence of multicollinearity. 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. For our purpose we only employ the ridge regression concept due to Hoerl and Kennard [17], to combat multicollinearity. There are a lot of works adopting ridge regression methodology to overcome the multicollinearity problem. To mention a few recent researches in full-parametric regression, see [36,15,20,21,6,5]. However Akdenïz and Tabakan [2], Roozbeh et al. [31], Akdenïz Duran et al. [3], Roozbeh and Arashi [30], Amini and Roozbeh [4], Arashi and Valizadeh [7] and Roozbeh [29] adopted this approach in facing with semiparametric regression model. The main focus of this approach is to develop necessary tools for computing the risk function of regression coefficient in a semiparametric regression model based on the eigenvalues of design matrix and then, estimating it based on robust approach.

The restricted models are widely applicable in the problem of general hypothesis testing specially the generalized likelihood ratio (GLR) tests in regression models. Norouzirad et al. [28] defined a restricted LASSO estimator and configure three classes of LASSO type estimators to fulfill both variable selection and restricted estimation in regression model. Akdenïz and Tabakan [2] and Akdenïz et al. [1] developed the restricted ridge and Liu estimators in semiparametric regression models. The problem of restricted ridge partial residual estimation in a semiparametric regression model with correlated errors is studied by Amini and Roozbeh [4] used generalized cross-validation (GCV) criteria for optimal bandwidth and ridge parameter selection in model (1.1), simultaneously.

Besides multicollinearity, outliers (points that fail to follow partial linear pattern of the majority of the points) are another common problem in the regression analysis. Robust regression methods are used to overcome the effects of outliers (inflated sum of squares, bias or distortion of estimation, distortion of p-values, etc.). Here, we only employ the least trimmed squares semiparametric regression estimators for both parts of our model. It is well-known that the ordinary least-squares estimator is very sensitive to outliers. This motivated the researchers to focus on robust estimators. Examples of recent researches in full-parametric regression include the studies made by Edelsbrunner and Souvaine [11], Bernholt [8], Jung [19], Erickson et al. [12], Bremner et al. [9], Nguyen and Welsch [27], Mount et al. [25,24] and Roozbeh and Babaie-Kafaki [32].

The basic measure of the robustness of an estimator is its breakdown point, that is, the fraction (up to 50%) of outlying data points that can corrupt the estimator arbitrarily. The study of efficient algorithms for robust statistical estimators have been an active area of research in computational geometry. Many researchers studied Rousseeuw's least median of squares estimator which is defined to be the hyperplane that minimizes the median squared residual (for example, see [33]). Although the vast majority of works on robust linear estimation in the field of computational geometry has been devoted to the study of the LMS estimator, it has been observed by Rousseeuw and Leroy [34] that LMS is not the estimator of choice from the perspective of statistical properties. They argued that a better choice is the least trimmed squares. The breakdown point of LTS and LMS are the same. Like LMS, the LTS estimator is a robust estimator with a 50%-breakdown point which means that the estimator is insensitive to the corruption made by outliers, provided that the outliers constitute less than 50% of the set. However, LTS has a number of advantages in contrast to LMS. The LTS objective function is smoother than that of LMS. LTS has better statistical efficiency because it is asymptotically normal (see [33]) and converges faster. Rousseeuw and van Driessen [35] remarked that, for these reasons, LTS is more suitable as a starting point for two-step robust estimators such as the MM-estimator (see [41]) and generalized M-estimators (see [37]).

The main focus of this paper is to study a robust generalized least squares ridge estimator in restricted semiparametric regression model. The organization of the paper is as follows: Section 2 contains the classical estimator of restricted semiparametric regression model based on kernel approach and related assumptions. The properties of generalized restricted ridge estimator of linear part are exactly derived in Section 3. We review least trimmed squares estimators in semiparametric regression model and then, propose a new robust estimator in restricted semiparametric regression model together with its theoretical properties in Section 4. We estimate the shrinkage parameter based on robust methods in Section 5 and then, the proposed generalized least squares ridge estimator will be reconstructed based on robust estimators of shrinkage parameter. In Section 6, the efficiencies of robust estimators relative to nonrobust estimator are evaluated for both ridge and nonridge types through the Monté-Carlo simulation studies as same as a real data example. Finally, some concluding important results are stated in Section 7.

2. The classical estimators under restriction

Consider the following semiparametric regression model

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