Statistics and Probability Letters 105 (2015) 106–113

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Influence diagnostic in ridge semiparametric models

Hadi Emami

Department of Statistics, University of Zanjan, Zanjan, Iran

ARTICLE INFO

Article history:
Received 17 December 2014
Received in revised form 9 June 2015
Accepted 9 June 2015
Available online 16 June 2015

Keywords:
Bandwidth
Cross-validation
Diagnostics
Ridge estimator
Smoothing spline

ABSTRACT

In this paper the ridge regression (RR) diagnostic method of Walker and Brich (1988) is applied to the ridge semiparametric regression model (RSPRM). We propose case deletion formulas to detect influential points. Furthermore, a real data set is analysed.

1. Introduction

Diagnostic techniques for the parametric regression model have received a great deal of attention in statistical literature since the seminal work of Cook (1977) and others including Cook and Weisberg (1982), Belsley et al. (1989) and Walker and Birch (1988). In semiparametric regression models (SPRMs), diagnostic results are quite rare; among them Eubank (1985), Thomas (1991) and Kim (1996) studied the basic diagnostic building blocks such as residuals and leverages, Kim et al. (2001, 2002) and Fung et al. (2002) proposed some type of Cook's distances in SPRMs.

It is not unusual to have influential cases and multicollinearity simultaneously in SPRMs. Frequently, the existence of influential observations is complicated by the presence of collinearity. However, there does not seem to be work on the possible effects that collinearity can have on the influence of an observation in SPRMs. In this paper, we therefore propose a case deletion formula to detect influential points in the ridge penalized least squares estimators (RPLSEs) of SPRM. We assess the global influence of observations on the RPLSEs using the method of case deletion suggested by Walker and Birch (1988). We also graphically show how the influence of a case can be modified when RPLSEs are used to reduce the level of collinearity. The paper is organized as follows. In the next section, SPRMs are introduced, the relevant notations and some inferential results are also given. In Section 3, the RPLSEs are defined based on the partial spline models. Section 4 derives some type of Cook's distance and case-deletion formulas for the RPLSEs. Statistical properties and motivation of these measures are also discussed. In Section 5 the proposed methods are illustrated through a real data set. A discussion is given in the last section.

2. Background and definition

Consider the SPRM given by

$$y_i = x_i'\beta + g(t_i) + \epsilon_i \quad 1 \le i \le n, \tag{1}$$

where β is a p-vector of regression coefficients, x_i is a p-vector of explanatory variables, t_i is a scalar ($a \le t_1, \ldots, t_n \le b$), and $t_i's$ are not all identical, g is a smooth function and the errors ϵ_i are uncorrelated with zero mean and constant variance σ^2.

E-mail address: h.emami@znu.ac.ir.

http://dx.doi.org/10.1016/j.spl.2015.06.010
0167-7152/

Model (1) has been used in discussion of many methods, e.g., penalized least squares (see Fung et al., 2002) and smoothing spline (see Green and Silverman, 1994). There are several ways of estimating β and g. Here the penalized least squares approach is of special interest. For this, let the ordered distinct values among $t_1, \ldots, t_n$ be denoted by $s_1, \ldots, s_q$. The connection between $t_1, \ldots, t_n$ and $s_1, \ldots, s_q$ is captured by means of $n \times q$ incidence matrix N, with entries $N_{ij} = 1$ if $t_i = s_j$ and 0 otherwise. Let g be the vector of value $a_i = g(s_i)$. For model (1) the penalized sum of squares is

$$\|y - X\beta - Ng\|^2 + \lambda \int g''(t)^2 dt, \tag{2}$$

where y is the vector of n response values and X is $n \times p$ design matrix. Minimizing (2) with respect to β and g, the partial spline least squares estimators (PLSEs) of β and g are

$$\hat{\beta} = \{X'(I_n - S)X\}^{-1}X'(I_n - S)y \tag{3}$$

and

$$\hat{g} = (N'N + \lambda M)^{-1}N'(y - X\hat{\beta}), \tag{4}$$

where I_n is the identity matrix of size n, $S = N(N'N + \lambda M)^{-1}N'$, λ is a nonnegative tuning parameter and M is a $q \times q$ matrix whose entries only depend on the knots $\{s_j\}$ (see Speckman, 1988 and Nishisato et al., 2002).

3. Partial spline models and ridge estimation

The procedure of fitting the model (1), essentially involves estimating the parameters of the model by assuming that $rank(X) = p$, or equivalently $rank\{X'(I_n - S)X\} = p$. In fact, if X is an ill-conditioned matrix, then the results may not fulfil our wishes, or can even be false in some situations, especially for small samples. There are a few studies that looked at overcoming the rank-deficient and ill-conditioned or multicollinearity problems in SPRMs (see Hu, 2005, Duran et al., 2012 and Roozbeh, 2015). Here, we use the RPLSEs which can be obtained by minimizing (2) subject to $\beta'\beta = d$, where d is constant. Particularly we should minimize

$$\|y - X\beta - Ng\|^2 + \lambda \int g''(t)^2 dt - k(\beta'\beta - d), \tag{5}$$

where k is a Lagrangian multiplier. Minimization of (5) can be done in a two steps estimation process: first we minimize it subject to $g(s_j) = a_j$, $j = 1, \ldots, q$ and in the second step we minimize the result over the choice of g and β. The problem of minimizing $\int g''(t)^2 dt$ subject to g interpolating given points $g(s_j) = a_j$ is given by Green and Silverman (1994), and minimizing function g provides a cubic spline with knots $\{s_j\}$. There exists a matrix M only depending on the knots $\{s_j\}$, such that the minimized value of $\int g''(t)^2 dt$ is $g'Mg$ (cf. Green and Silverman, 1994, p. 66). The equation in (5) is therefore of the form

$$\|y - X\beta - Ng\|^2 + \lambda g'Mg - k(\beta'\beta - d). \tag{6}$$

Now, by minimizing (6) the RPLSEs of β and g are obtained as

$$\hat{\beta}_k = \{X'(I_n - S)X + kI_p\}^{-1}X'(I_n - S)y \tag{7}$$

and

$$\hat{g} = (N'N + \lambda M)^{-1}N'(y - X\hat{\beta}_k), \tag{8}$$

where I_p is identity matrix of size p. In this paper, the choice of the smoothing parameter λ is accomplished by minimizing the generalized cross-validation criterion $GCV(\lambda)$. In the literature k has been referred as ridge parameter. It is obvious that for k equal to zero, the RPLSEs defined in (7) and (8) are exactly the same as PLSEs of β and g defined in (3) and (4). The choice of this parameter in ordinary least squares is still unsolved and because of this problem several approaches have been developed to guide data analysts in the selection of shrinkage parameter (see Walker and Birch, 1988). A common approach suggests plotting scheme mean squared error on element of ridge estimator versus different values of k and choosing the best value for k. Here we follow the method of Hu (2005) to choose the balance parameter k. At first, we choose some $\{k_l, l = 1, 2, \ldots, h\}$ such that $rank\{X'(I_n - S)X + kI_p\} = p$, and then we find a k_{l_0} from $\{k_l, l = 1, 2, \ldots, h\}$ such that k_{l_0} minimizes the term in (2).

4. Influence diagnostic in RPLSE

4.1. Leverage and residuals

From (7) and (8) the vector of fitted values can be written as

$$\begin{aligned}\hat{y} &= X\hat{\beta}_{(k)} + N\hat{g} \\ &= (\tilde{H}_k + H_k^*)y \\ &= H_k y,\end{aligned}$$