Volume-4, Issue-10, Oct-2015 • ISSN No 2277 - 8160

Research Paper Statistics

Analysis of Influential Observation for Ridge Estimator in Semiparametric Regression

Semra TÜRKAN Hacettepe University, Department of Statistics, Beytepe, 06800Ankara, Turkey

ABSTRACT *The detection of infuential observations has attracted a great deal of attention in last few decades. Most of the ideas of determining infuential observations are based on single-case diagnostics with ith case deleted. The Cook's distance is most commonly used among the other single-case diagnostics and successfully applied to various statistical models. In this article, we propose Cook's distance for the ridge regression estimator of the parametric component in the semiparametric regression model to detect infuential observations. We investigate the performance of proposed diagnostic to detect infuential observations by using simulation data.*

KEYWORDS : Semiparametric regression model, Ridge regression estimator, Cook's distance, Infuential observations.

1. Introduction

Regression diagnostics consist of a collection of methods used in the identification of influential points and multicollinearity. Particularly, the detection of influential observations has received a great deal of attention in the statistical literature. Hence a number of statistical measures have been proposed to identify influential observations, in which Cook's distance proposed by Cook (1977) is one of the most commonly used influence measures. However, most of the regression diagnostics have been related to parametric regression models. Diagnostic measures in various nonparametric regression or semiparametric regression models are quite rare (Zhang et al., 2007).

In this paper, we examine the influence of observations on the ridge estimator for the vector of parameters $\boldsymbol{\beta}$ in a semiparametric regression model. Consider the following semiparametric regression model:

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + f(t_i) + \varepsilon_i, \qquad i = 1, 2, ..., n \tag{1}$$

where the y_i's are observations, $\mathbf{x}_i$ is px1 vector ($p \le n$), t_i is scalar, $\boldsymbol{\beta} = (\beta_1,...,\beta_p)^T$ is a vector of unknown parameters, f(.) is unknown function, and ε_i 's are independent and identically distributed random variables with zero mean and variance σ^2.

Semiparametric models are more flexible than linear models since they consist of a parametric and a nonparametric component. These models are used when the response y linearly depends on x, but it is nonlinearly related to t. Hence f(t) represents a smooth unparametrized functional relationship. The aim is to estimate $\boldsymbol{\beta}$ and nonparametric function f(t) from the data $\{y_i, x_i, t_i\}$.

The model in (1) can be written in matrix-vector notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon} \tag{2}$$

where $\mathbf{y} = (y_1,..., y_n)^T$, $\mathbf{f} = (f(t_1),...,f(t_n))^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1,...,\varepsilon_n)$ and $\mathbf{X} = (\mathbf{x}_1,...,\mathbf{x}_n)^T$ is the n x p

matrix without any restrained conditions, namely, *rank (X) <p* or *rank (X) = p* (ill-conditioned or not).

In fact, if ***X*** is an ill-conditioned matrix, then the results may not reliable. Many studies do not consider the case *rank (X)<p*, and the situation that the design matrix ***X*** is rank-deficient is rarely investigated. It is noticeable that ridge estimation not only solves rank-deficient and ill-conditioned problems, but also presents a new method which can deal with (non)linear and semiparametric regression models for *rank(X)=p* without ill-conditioning (Hu 2005). Hence, Hu (2005) proposed ridge estimator of semiparametric regression model. Roozbeh et al. (2010) introduced a semiparametric ridge regression estimator for the vector-parameter when the matrix $\mathbf{X}^T\mathbf{X}$ is ill-conditioned. The large condition number of parametric component indicates that a ridge regression estimator can be used for $\boldsymbol{\beta}$. The reduction in multicollinearity should be a first step for the effective detection of influential observations.

In this article, the Cook's distance is defined to detect influential observations on the ridge estimator of $\boldsymbol{\beta}$ in semiparametric regression. Approximate deletion formula as functions of corresponding residuals and leverages are proposed. In Section 2, the ridge regression estimator is given. In Section 3, the approximate deletion formula of Cook's distance for the ridge estimator of $\boldsymbol{\beta}$ in semiparametric regression is derived. In section 4, the simulation data is examined using proposed diagnostic method.

The Model and Ridge Estimator

In this article, the partial kernel smoothing estimator of $\boldsymbol{\beta}$, which attains the usual parametric convergence rate $n^{1/2}$ without undersmoothing the nonparametric component f(.), is used. If $\boldsymbol{\beta}$ is known, a naturel nonparametric estimator of f(.) is

$$\hat{f}(t,\boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ni}(t)(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) \qquad (3)$$

where $W_{ni}(.)$ is the positive weight function (Roozbeh et al., 2010).

To estimate $\boldsymbol{\beta}$ in (2) using kernel weight functions, least square estimator can be used. Consider the following objective function:

$$SS(\boldsymbol{\beta}) = (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) \qquad (4)$$

where $\tilde{\mathbf{y}} = (\tilde{y}_1,...,\tilde{y}_n)$, $\tilde{\mathbf{X}} = (\mathbf{x}_1^T,...,\mathbf{x}_n^T)$, $\tilde{y}_i = y_i - \sum_{j=1}^{n} W_{nj}(t_i)y_j$ and $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \sum_{j=1}^{n} W_{nj}(t_i)\mathbf{x}_j$ for i=1,…,n. The first-order condition of objective function (4) minimized by the vector $\boldsymbol{\beta}$ is obtained as:

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{y}} \qquad (5)$$

The properties of the least square estimation of $\boldsymbol{\beta}$ depend heavily on the characteristics of the $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$. If the $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ matrix is ill-contitioned (close dependency among various columns of $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$), the least square estimation produce unduly large sampling variances. As a remedy, Hu

(2005) suggested to use ridge regression estimator instead of least square estimator. The ridge regression estimator of $\boldsymbol{\beta}$ can be obtained by using following objective function:

$$SS(\boldsymbol{\beta}) = (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}) + k\boldsymbol{\beta}^T\boldsymbol{\beta} \quad (6)$$

The first-order condition of objective function (6) minimized by the vector $\boldsymbol{\beta}$ is obtained as:

$$\hat{\boldsymbol{\beta}}_R = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + k\mathbf{I})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{y}} \quad (7)$$

where k is a biasing parameter $(k > 0)$ and $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{W})\mathbf{X}$, $\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{W})\mathbf{y}$. For the model in (2), the estimates of $\boldsymbol{\beta}$ and $\mathbf{f}$ can be written as

$$\hat{\boldsymbol{\beta}}_R = \mathbf{A}\mathbf{y} \quad (8)$$

$$\hat{\mathbf{f}}_R = \mathbf{W}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R) \quad (9)$$

where $\mathbf{A} = \left[\mathbf{X}^T(\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})\mathbf{X} + k\mathbf{I}\right]^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$ and $\mathbf{W}$ is kernel weights matrix (Tabakan and Akdeniz, 2009).

Cook's Distance for Ridge Estimator

In this section, we define Cook's distance for ridge estimator of $\boldsymbol{\beta}$ in semiparametric regression to gauge the influential observations and express it as function of the corresponding residuals and leverages.

3.1. Influence on $\hat{\boldsymbol{\beta}}$

Following the study of Walker and Birch (1988), an influence measure for the *i*th observation on $\hat{\boldsymbol{\beta}}$ can be defined as a type of Cook's distance for ridge regression

$$\tilde{C}_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{R,-i})^T\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{R,-i})}{\sigma^2\,\mathrm{tr}(\tilde{\mathbf{H}})} \quad (10)$$

where $\tilde{\mathbf{H}} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + k\mathbf{I})^{-1}\tilde{\mathbf{X}}^T$ is matrix which plays the same role as the hat matrix in least square regression. Noting $\mathrm{tr}(\tilde{\mathbf{H}}) = p$, the above defined $\tilde{C}_i$ can be expressed as a function of the corresponding residuals and leverages,

$$\tilde{C}_i = \frac{\tilde{e}_i^2}{p\sigma^2}\left[\frac{\sum_{i=1}^{n}\tilde{h}_{ij}^2}{(1-\tilde{h}_i)^2}\right] \quad (11)$$

where $\tilde{e}_i$ is the i*th* component of the residual vector $\tilde{\mathbf{e}} = \tilde{\mathbf{y}} - \hat{\mathbf{y}}_R = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_R$, $\tilde{h}_i$ is i*th* diagonal component of $\tilde{\mathbf{H}}$ and $\tilde{h}_{ij}$ is the ij*th* element of $\tilde{\mathbf{H}}$. To compute $\tilde{C}_i$, it is required an estimator

of σ^2. A simple estimator is $s^2 = \sum_{i=1}^{n} \tilde{e}_i^2 /(n-p)$.

3.2. Influence on $\hat{y}$

An influence measure for the *i*th observation on the vector of fitted values can be similarly defined by

$$C_i = \frac{(\hat{y}_i - \hat{y}_{R,-i})}{SE(\hat{y}_i)}$$
$$= \frac{\tilde{x}_i^T(\hat{\beta}_R - \hat{\beta}_{R,-i})}{SE(\hat{y}_i)} \qquad (12)$$

where $SE(\hat{y}_i) = s\left[\mathbf{x}_i^T(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + k\mathbf{I})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + k\mathbf{I})^{-1}\mathbf{x}_i\right]^{1/2}$ is an estimator of standart error of the fitted value.

C_i can be expressed as a function of the corresponding residuals and leverages as

$$C_i = \frac{1}{s(\sum_{i=1}^{n}\tilde{h}_{ij}^2)^{1/2}} \frac{\tilde{h}_i\tilde{e}_i}{(1-\tilde{h}_i)} \qquad (13)$$

The main advantage of deletion formulas is that, as in LS, the estimator does not have to be recomputed every time a case is deleted. For a value of k and h, all of the elements of (11) and (13) are readily available from a single run of semiparametric regression.

Simulation Study

In this section, we consider the model generating the data set which is used in the study of Roozbeh et al. (2010). However, we generate the data that will be included influential observations. The following model is used to generate data:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4 + x_{i5}\beta_5 + f(t_i) + \varepsilon_i, \qquad i = 1,...,n$$

where $\boldsymbol{\beta} = (1,2,2,-5,4)$, $\sigma = 0.1$ and $f(t_i) = \sqrt{t_i(1-t_i)}\sin(\frac{2.1\pi}{t_i + 0.05})$ for $t_i = (i-0.5)/n$, i=1,…,n. The first 450 observations are generated from $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and $\mathbf{x}_i \sim N_5(\mathbf{0}, \sum_x)$ with

$$\sum_x = \begin{pmatrix} 0.81 & 0.4 & 0.3 & -0.2 & -0.1 \\ 0.4 & 2.25 & 0.4 & 0.3 & -0.2 \\ 0.3 & 0.4 & 1 & 0.4 & 0.3 \\ -0.2 & 0.3 & 0.4 & 0.64 & 0.4 \\ -0.1 & -0.2 & 0.3 & 0.4 & 0.49 \end{pmatrix}$$

and the last 50 observations which are considered as influential are generated from $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ where $\sigma = 2$ and $\mathbf{x}_i \sim N_5(\boldsymbol{\mu}, \sum_x)$ with $\boldsymbol{\mu} = (10,10,10,10,10)$ and $\sum_x$.

The weight function $W_{ni}(t_j)$ is defined as

$$W_{ni}(t_j) = \frac{1}{nh_n} K\left(\frac{t_i - t_j}{h_n}\right) = \frac{1}{nh_n}\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(t_i - t_j)}{2h_n^2}\right\}$$

which is Priestley and Chao's weight with the Gaussian kernel (Roozbeh et al., 2010).
We use the cross-validation (CV) method to select the optimal bandwith h_n and biasing parameter k simultaneously, which minimizes the following CV function:

$$CV(h_n) = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \tilde{y}_{-i})^2 = \frac{1}{n}\sum_{i=1}^{n}(\tilde{e}_i / 1 - \tilde{h}_{ii})^2 ,$$

Different combinations of k (0.2, 0.4, 0.6, 0.8, 1) and h_n= (0.01, 0.06, 0.11, 0.16, 0.21, 0.26, …, 0.96) are used to find k and h_n which minimize the CV simultaneously. The minimum of CV occured at k=0.2 and h_n =0.06.

The ratio of largest eigenvalue to the smallest eigenvalue of matrix $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ is $\lambda_5 / \lambda_1 = 29571.84/38.12 = 775.75$ which implies that $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ matrix is ill-contitioned or the existence of multicollinearity in the data set. Hence, in this situation it is supported that ridge estimator is used instead of least square estimates.

In order to examine the efficiency of Cook's distance $\tilde{C}_i$ defined in (10) and C_i defined in (12) in detecting single observation that has large impact on ridge estimator of $\boldsymbol{\beta}$ in semiparametric regression, as stated above, some observations are generated to be influential obsevations one by one to see whether the defined measures detect it out. For each observation, these measures are calculated and the index plots of these measures are obtained as in Figure 1 and Figure 2.

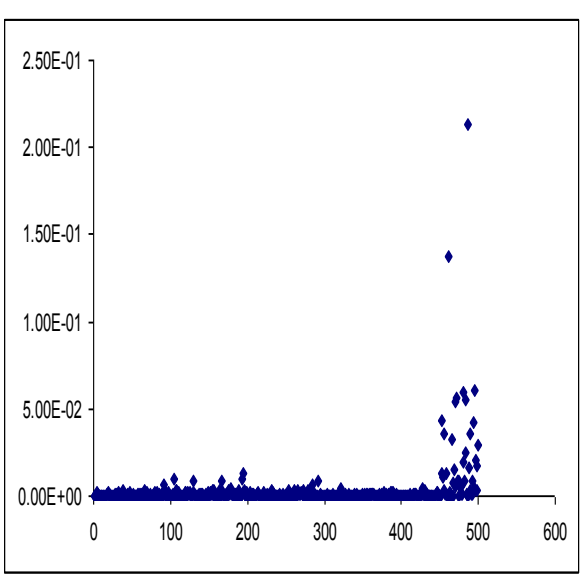

Figure 1. Index plot of $\tilde{C}_i$

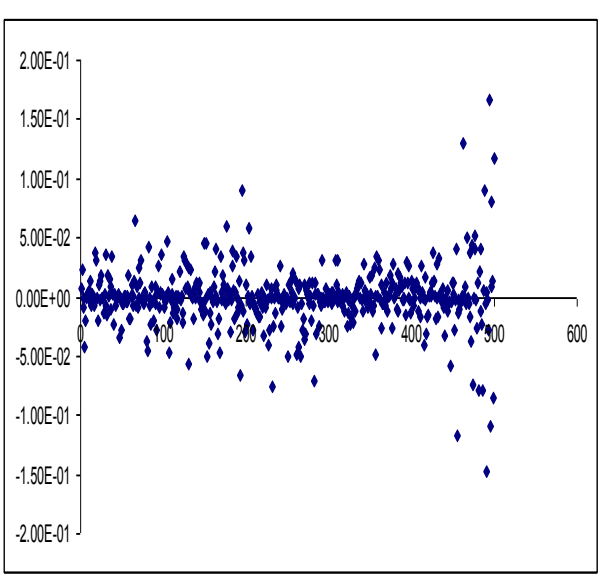

Figure 2. Index plot of C_i

As seen from Figure 1 and Figure 2, $\widetilde{C}_i$ can recognize majority of generated the last 50 influential observations while C_i can not as successful as $\widetilde{C}_i$ to recognize generated influential observations. In addition, C_i could find some other observations as influential observations.

Conclusion

In this study, Cook's distances that are function of leverages and residuals is studied for ridge estimator in semiparametric regression. Although no conventional cut-off points are introduced or developed for the these measures, it seems that index plot is an optimistic and conventional procedure to disclose infuential cases. It is seen that the proposed Cook's distances are successful to detect influential observations in the data.

REFERENCES

Akdeniz, F., Tabakan, G.: Restricted Ridge Estimators of the Parameters in Semiparametric Regression Model, Communications in Statistics-Theory and Methods, 38: 1852-1869 (2009). Cook, R.D.: Detection of Influential Observations in Linear Regression, Technometrics, 19, 15-18 (1977). Hu, H.:Ridge estimation of semiparametric regression model. J. Comput. Appl. Math. 176, 215-222 (2005). Roozbeh, M., Arashi, M., Niroumand, H.A.: Semiparametric ridge regression approach in partially linear models. Comm. Statist. Simulation Comput. 39, 449-460 (2010). Zhang, C., Mei, C., Zhang, J.: Influence Diagnostics in Partially Varying Coefficient Models, Acta. Math. Appl. Sinica, 23 (4): 619-628 (2007). Walker, E. and Birch, J.B. Inuence measures in ridge regression. Technometrics 30, 221-227 (1988).