Revista Colombiana de Estadística
Diciembre 2013, volumen 36, no. 2, pp. 273 a 286

Detection of Influential Observations in Semiparametric Regression Model

Detección de observaciones influenciales en modelos de regresión semiparamétricos

Semra Türkan[a], Öniz Toktamis[b]

Department of Statistics, The Faculty of Science, Hacettepe University, Ankara, Turkey

Resumen

In this article, we consider the semiparametric regression model and examine influential observations which have undue effects on the estimators for this model. One of the approaches to measure the influence of an individual observation is to delete the observation from the data. The most common measure based on this approach is Cook's distance. Recently, Daniel Peña introduced a new measure based on this approach. Pena's measure is able to detect high leverage outliers, which could be undetected by Cook's distance, in large data sets in linear regression model. The Cook's distances for parameter vector, unknown smooth function and response variable in semiparametric regression model are expressed by authors as functions of the residuals and leverages. Following the study of them we derive a type of Pena's measure as functions of the residuals and leverages for the same model. We compare the performance of these measures as to detection of influential observations using real data, artificial data and simulation. The results show that the performance of Pena's measure is better than Cook's distance to detect high leverage outliers in large data sets in the semiparametric regression model such as in the linear regression model.

***Palabras clave*:** Cook's distance, High leverage outliers, Pena's measure, Semiparametric regression.

Abstract

En este articulo, se consideran modelos de regresión semiparamétrica y se examinan observaciones influenciales que pueden tener efectos sobre los estimadores para este modelo. Una de las formas de medir la influencia de una observación individual es borrando la observación en el conjunto de

[a]Doctor. E-mail: sturkan@hacettepe.edu.tr
[b]Emeritus professor. E-mail: oniz@hacettepe.edu.tr

datos. La medida más común bajo esta idea es la distancia de Cook. Recientemente, Daniel Peña introdujo una nueva medida basada en estas ideas. Las distancias de Cook para el vector de parámetros, la función de suavizamiento y la variable respuesta en modelos de regresión semiparamétrica han sido expresadas por otros autores como funciones de los residuales y los puntos de apalancamiento. Se deriva en este artículo, una medida del tipo de la de Peña como función de los residuales y puntos de apalancamiento para el mismo modelo. Se compara el desempeño de estas medidas para la detección de observaciones influenciales usando datos reales y bajo simulación. Los resultados muestran que la medida de Peña es mejor que la distancia de Cook para detectar outliers y puntos de apalancamiento en conjuntos de datos grandes en los modelos de regresión semiparamétrica tales como el modelo de regresión lineal.

***Key words*:** distancia de Cook, outliers, puntos de apalancamiento, medida de Peña, regresión semiparamétrica.

1. Introduction

One or few observations could have serious effects on estimators. When an observation is omitted from the analysis, the fitted equation may change hardly at all. In this situation, the observation is considered as an influential observation. Hence, the detection of these observations has received a great deal of attention in the last decades. Numerous influence measures have been developed to detect these observations. Firstly, Cook (1977) introduced Cook's distance, which is based on deleting the observations one after another and measuring their effects in linear regression. Following the study of Cook (1977), most of ideas of detecting influential observations based on the deleting approach have developed. In recent years, Pena's measure is one of these ideas.

The study of influential observations has been extended to other statistical models using similar ideas such as in linear regression. However, most of the influence measures are concerned with parametric regression models. In recent years, the detection of influential observations in the nonparametric regression and semiparametric regression have been studied (see Thomas 1991, Kim 1996, Kim & Kim 1998, Kim, Park & Kim 2001, Zhu & Wei 2001, Kim, Park & Kim 2002, Zhang, Mei & Zhang 2007).

In this article, we consider the influence of individual cases on estimators in the semiparametric regression model and adjust the Pena's measure (Pena 2005) for this model. We compare the Pena's measure and some types of Cook's distances suggested by Kim et al. (2002) as to the success of detection of high leverages outliers in the semiparametric regression model.

The study is organized as follows. In Section 2, the semiparametric regression model is introduced. In Section 3, the formulas of Cook's distances for semiparametric regression model are given. In Section 4, Pena's measure formula for semiparametric regression is derived. In Section 5, the success of these measures

to detect influential observations, particularly high leverages outliers in large data, is analyzed via real data, artificial data and simulation.

2. Semiparametric Regression

Consider a semiparametric regression model with k explanatory variables

$$y_i = \mathbf{z_i}^T\boldsymbol{\beta} + m(x_i) + \varepsilon_i,\ (1 \leq i \leq n)$$

where y_i's are outcomes, $\mathbf{z_i}$ is a $k \times 1$ vector related to parametric component, x_i is a scalar, $\boldsymbol{\beta}$ is the $k \times 1$ vector of unknown parameters and m is a smooth unknown function. There are many approaches to estimate $\boldsymbol{\beta}$ and $\boldsymbol{m}$. The Speckman approach is one of them. Here, we follow the Speckman approach.

Let $\widetilde{\mathbf{Z}} = (\mathbf{I} - \mathbf{S})\mathbf{Z}$ and $\widetilde{\mathbf{y}} = (\mathbf{I} - \mathbf{S})\mathbf{y}$ where $\mathbf{S}$ is a smoother matrix. The local polynomial and the spline estimators are two classes of smoothers in semiparametric regression. Here, we use a local polynomial estimator. Hence, the $(1 \times n)$ jth row vector of $\mathbf{S}$ could be defined as $\mathbf{S}_{xj} = \mathbf{t}^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x$ where $\mathbf{X}_x$ is the $n \times (p+1)$ matrix with its ijth element equal to $(x_i - x)^{j-1}$, $\mathbf{W}_x = Diag(K_h(x_i - x))$ is the weight matrix with $K_h(.) = K(.\,|h)/h$ being a kernel function and h bandwidth controlling the size of the local neighborhood and $\mathbf{t}^T = \mathbf{t}_x^T(x) = (1, x - x, \ldots, (x - x)^p)$ is a vector. Here, it is assumed that K is a symmetric probability density function. The estimators of $\boldsymbol{\beta}$ and $\mathbf{m}$ suggested in Speckman (1988) are given by

$$\widehat{\boldsymbol{\beta}} = \left(\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}}\right)^{-1}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{y}} \tag{1}$$

$$\widehat{\mathbf{m}}(x) = \mathbf{S}\left(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\beta}}\right) = \mathbf{S}(\mathbf{I} - \widehat{\mathbf{H}})\mathbf{y} = \mathbf{H}^*\mathbf{y} \tag{2}$$

where $\widehat{\mathbf{H}} = (\mathbf{I} - \mathbf{S})^{-1}\widetilde{\mathbf{Z}}\left(\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}}\right)^{-1}\widetilde{\mathbf{Z}}^T(\mathbf{I} - \mathbf{S})$ and $\mathbf{H}^* = \mathbf{S}\left(\mathbf{I} - \widehat{\mathbf{H}}\right)$. The vector of fitted values could be expressed from (1) and (2) as below

$$\begin{aligned}\widehat{\mathbf{y}} &= \mathbf{Z}\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{m}}(x)\\ &= \breve{\mathbf{H}}\mathbf{y}\end{aligned} \tag{3}$$

where $\breve{\mathbf{H}}$ is considered as hat matrix in linear regression model defined $\breve{\mathbf{H}} = \widehat{\mathbf{H}} + \mathbf{H}^*$. The residual vector is given by

$$\breve{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \breve{\mathbf{H}})\mathbf{y}$$

which will be used in defining and interpreting Cook's distances in the semiparametric regression model.

3. Cook's Distance

Firstly, we briefly review the derivation of Cook's distance in the linear regression model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is a response vector, $\mathbf{X}$ is a $n \times k$ matrix of

known covariates, $\boldsymbol{\beta}$ is a vector of unknown parameters, and $\boldsymbol{\varepsilon}$ is a vector of errors with mean zero and a common unknown variance σ^2. y_i and $\mathbf{x}_i^T$ denote the ith row of $\mathbf{y}$ and $\mathbf{X}$, respectively, and using the subscript $(-i)$ means that the ith observation is deleted. Hence, $\mathbf{X}_{-i}$ denotes the matrix $\mathbf{X}$ with$i$th row deleted. Let $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ be the least squares estimator of $\boldsymbol{\beta}$, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the hat matrix and $s^2 = \mathbf{e}^T\mathbf{e}/(n-k)$ is estimation of σ^2.

Cook's distance for measuring the influence of the ith observation is defined by

$$C_i = (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-i})^T(\mathbf{X}^T\mathbf{X})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-i})/s^2 tr(\mathbf{H})$$

Using the fact,

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-i} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i e_i/(1-h_{ii})$$

the Cook's distance can be written as leverage values and residuals

$$C_i = \frac{1}{tr(\mathbf{H})s^2}\frac{e_i^2 h_{ii}}{(1-h_{ii}^2)} \tag{4}$$

where h_{ii} is the diagonal elements of $\mathbf{H}$ and e_i is the element of residual vector $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$. The trace of $\mathbf{H}$ is defined to be the sum of the elements on the main diagonal of $\mathbf{H}$. As a projection matrix, $\mathbf{H}$ is symmetric and idempotent ($\mathbf{H}^2 = \mathbf{H}$), the eigenvalues of a projection matrix are either zero or one and the number of non zero eigenvalues is equal to the rank of the matrix. In this case, $rank(\mathbf{H}) = rank(\mathbf{X}) = k$ and hence, $trace(\mathbf{H}) = k$ which means that $tr(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = k$.

3.1. Cook's Distance for $\widehat{\beta}$ in Semiparametric Regression

An influence measure for ith observation on $\widehat{\boldsymbol{\beta}}$ may be defined as a type of Cook's distance in linear regression by

$$\widetilde{C_i} = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-i})^T(\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-i})}{s^2 tr(\widetilde{\mathbf{H}})} \tag{5}$$

Note that $tr(\widetilde{\mathbf{H}}) = \sum_{i=1}^{n} \widetilde{h}_{ii} = k$ as in linear regression. Equation (5) can be expressed as a function of the ith residual and leverage such as in (4) for semiparametric regression model as below

$$\widetilde{C}_i = \frac{1}{s^2 k}\frac{\widetilde{h}_{ii}\widetilde{e}_i^2}{(1-\widetilde{h}_{ii})^2} \tag{6}$$

where $\widetilde{e}_i$ is the ith component of residual vector $\widetilde{\mathbf{e}} = \mathbf{y} - \widetilde{\mathbf{y}}$ and $\widetilde{h}_{ii}$ is the ith diagonal component of $\widetilde{\mathbf{H}} = \widetilde{\mathbf{Z}}(\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}^T$ related to parametric component of semiparametric regression model (Kim et al. 2002).

3.2. Cook's Distance for $\widehat{\mathbf{m}}$ in Semiparametric Regression

An influence measure for ith observation on $\widehat{\mathbf{m}}$ may be defined as a type of Cook's distance utilizing (2) by

$$C_i^* = \frac{\{\widehat{m}(x_i) - \widehat{m}_{-i}(x_i)\}}{s^2 tr(\mathbf{H}^*)}$$

It can be expressed as a function of the ith residual and leverage such as in (4)

$$C_i^* = \frac{(h_{ii}^* e_i^*)^2}{(1 - h_{ii}^*)^2 s^2 tr(\mathbf{H}^*)} \tag{7}$$

where e_i^* is the ith component of residual vector $\mathbf{e}^* = (\mathbf{I} - \mathbf{H}^*)\mathbf{y}$ and h_{ii}^* is the ith diagonal component of $\mathbf{H}^*$ related to the nonparametric component of the semiparametric regression model (Kim et al. 2002).

3.3. Cook's Distance for $\widehat{\mathbf{y}}$ in Semiparametric Regression

An influence measure for ith observation on $\widehat{\mathbf{y}}$ may be defined as a type of Cook's distance utilizing (3) such as in linear regression by

$$\breve{C}_i = \frac{(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{-i})^T (\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{-i})}{s^2 tr(\breve{\mathbf{H}})}$$

It can be expressed as a function of the ith residual and leverage such as in (4) for $\widehat{y}$

$$\breve{C}_i = \frac{\breve{h}_{ii} \breve{e}_i^2}{(1 - \breve{h}_{ii})^2 s^2 tr(\breve{\mathbf{H}})} \tag{8}$$

where $\breve{e}_i$ is the ith component of residual vector $\breve{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \breve{\mathbf{H}})\mathbf{y}$ and $\breve{h}_{ii}$ is the ith diagonal component of $\breve{\mathbf{H}}$ (Kim et al. 2002).

4. Pena's Measure

Pena (2005) introduced a new measure to determine the influence of an observation based on how this observation is being influenced by the rest of the data. That is, the predicted change when each observation in the data is deleted is measured for each observation. In this way, the sensitivity of each observation to changes in the data is measured. Pena (2005) showed that this type of influential analysis is able to indicate features in the data, such as clusters of high leverage outliers. Pena's measure has some advantages over Cook's distance. In a sample without outliers or high leverage observations, all of the cases have the the same expected sensitivity with respect to the entire sample. This is an advantage over Cook's distance which has an expected value that depends heavily

on the leverage of the case. For large sample sizes with many predictors, the distribution of the Pena's measure will be approximately normal. This is advantage over Cook's distance which has a complicated asymptotical distribution. The sample contaminated by a group of similar outliers with high leverages, this measure could discriminate between outliers and good observations while Cook's distance fails to detect these observations. In addition, Pena's measure can be useful for identifying intermediate-leverage outliers that are not detected by Cook's distance (Pena 2005).

In the regression model, Pena's measure is defined as

$$S_i = \frac{\mathbf{s}_i^T \mathbf{s}_i}{p s^2_{\widehat{(y_i)}}} \tag{9}$$

where $\mathbf{s}_i = (\widehat{y}_i - \widehat{y}_{i(1)}, \ldots, \widehat{y}_i - \widehat{y}_{i(n)})$ is a vector and $\widehat{y}_{i(j)}$ is the ith fitted value when the jth observation is deleted. Using the facts, the difference $\widehat{y}_i - \widehat{y}_{i(j)}$ is obtained as

$$\widehat{y}_i - \widehat{y}_{i(j)} = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{-j} = \frac{h_{jj} e_j}{1 - h_{jj}} \text{ and } s^2_{\widehat{(y_i)}} = s^2 h_{ii} \tag{10}$$

Pena's measure can be expressed as a function of the ith residual and leverage from (10)

$$S_i = \frac{1}{p s^2 h_{ii}} = \sum_{j=1}^{n} \frac{h_{ji}^2 e_j^2}{(1 - h_{jj})^2} \tag{11}$$

Pena (2005) stated that S_i would be large if it exceeds median (S_i)+$4.5MAD(S_i)$ where $MAD(S_i) = median\{|S_i - median(S_i)|\}/0,6745$. Pena's measure is very effective in detection of high leverage outliers that can not be detected by Cook's distance in large data sets. Also, it is very simple to compute (Türkan, S. and Toktamis, Ö. 2012).

4.1. Pena's Measure for Semiparametric Regression

In this study, we derived Pena's measure formula for the semiparametric regression model. The fitted values vector in (3) can be written as

$$\begin{aligned} \widehat{\mathbf{y}} &= \mathbf{Z}\boldsymbol{\beta} + \widehat{\mathbf{m}}(x) \\ &= \widetilde{\mathbf{Z}}\widehat{\boldsymbol{\beta}} + \mathbf{S}\mathbf{y} \end{aligned} \tag{12}$$

Using ith row vector of $\mathbf{S}$ in (12), $\mathbf{S}_{xi} = \mathbf{t}^T(\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x$, the ith fitted value, $\widehat{y}_i$, can be written

$$\widehat{y}_i = \widetilde{\mathbf{z}}_i^T \widehat{\boldsymbol{\beta}} + \mathbf{t}_{x_i}(x_i) \widehat{\boldsymbol{\beta}}_{x_i}$$

where $\widehat{\boldsymbol{\beta}}_x = (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{y}$ and $t_x(x_i) = (1, (x_i - x), \ldots, (x_i - x)^p)$. The ith fitted value when jth observation is deleted, $\widehat{y}_{i,-j}$, can be expressed as below:

$$\widehat{y}_{i,-j} = \widetilde{\mathbf{z}}_i^T \widehat{\boldsymbol{\beta}}_{-j} + \mathbf{t}_{x_i}(x_i) \widehat{\boldsymbol{\beta}}_{x_i,-j} \tag{13}$$

Utilizing Sherman-Morrison-Woodbury (SMW) theorem, $\widehat{y}_i - \widehat{y}_{i,-j}$ can be obtained as a function of the ith residuals and leverages

$$\widehat{y}_i - \widehat{y}_{i,-j} = \frac{\widetilde{h}_{jj}\widetilde{e}_j}{1-\widetilde{h}_{jj}} + \frac{h_{x_i}(j,j)e_{x_i(j)}}{1-h_{x_i}(j,j)} \tag{14}$$

where $\widetilde{h}_{ij} = \widetilde{\mathbf{z}}_i^T(\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{z}}_j$ and $h_{x_i}(i,i) = (\mathbf{X}_{x_i}^T\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}K_h(0)$ are diagonal elements of $\widetilde{\mathbf{H}} = \widetilde{\mathbf{Z}}(\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}^T$ and $\mathbf{H}_x = \mathbf{X}_x(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x$, respectively. From (14), Pena's measure for semiparametric regression model can be obtained as

$$\begin{aligned}\widetilde{S}_i &= \frac{\mathbf{s}_i^T\mathbf{s}_i}{tr(\breve{\mathbf{H}})var(\widehat{y}_i)} \\ &= \frac{1}{tr(\breve{\mathbf{H}})var(\widehat{y}_i)}\sum_{j=1}^{n}\left(\frac{\widetilde{h}_{jj}\widetilde{e}_j}{1-\widetilde{h}_{jj}} + \frac{h_{x_i}(j,j)e_{x_i(j)}}{1-h_{x_i}(j,j)}\right)^2\end{aligned} \tag{15}$$

(see Türkan 2012)

5. Application

In this section, we compare the performance of our adjusted Pena's measure with adjusted Cook's distances in the semiparametric regression model to identify influential observations via actual data, artificial data and a simulation.

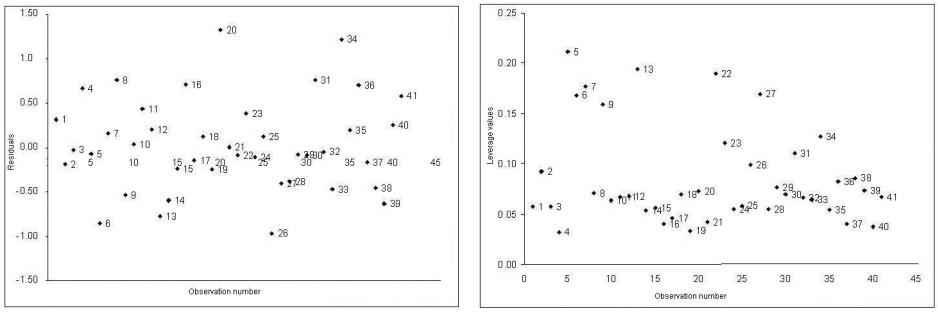
5.1. Actual Data

We consider actual data related to diabetes. The response variable is the logarithm of C-peptide concentration (y) at diagnosis and two predictors are age (x) and base deficit (z) (Kim et al. 2002). The data set contains 41 observations. There is a linear relationship between the logarithm of C-peptide concentration and base deficit, however, there is a nonlinear relationship between the logarithm of C-peptide concentration and age. Hence, the semiparametric regression model, $y_i = \mathbf{z}_i^T\boldsymbol{\beta} + m(x_i) + \varepsilon$, is used. Following the study of Kim et al. (2002), the local linear smoother was used and the bandwidth $h = 5.6$ was selected minimizing cross-validation (CV) criterion ($CV = \sum\{e_i/(1-h_{ii})\}^2$). Table 1 shows the estimates of both parametric and nonparametric components.

Figure 1 displays index plots of leverages values $\breve{h}_{ii}$ and residuals $\breve{e}_i$.

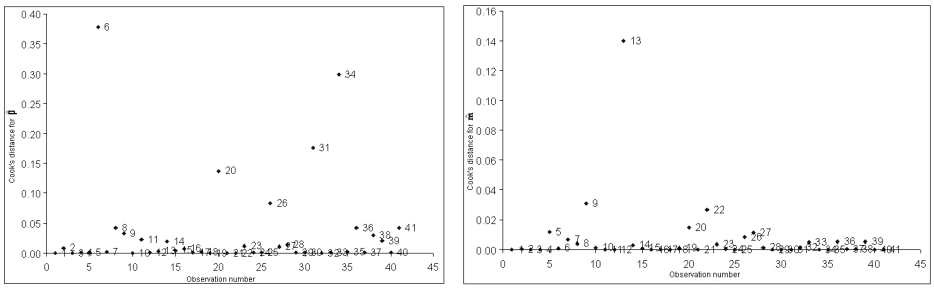
As seen from Figure 1(a), observations 20 and 34 are considered as outliers but these observations are not considered as high leverage from Figure 1(b) that the values of $\breve{h}_{ii}$ are not close to 1. Hence, it is said that there is no high leverage outlier in the data.

Figure 2 displays an index plot of influence measures ($\widetilde{C}, C_i^*, \breve{C}_i$ and $\widetilde{S}_i$) for this data.

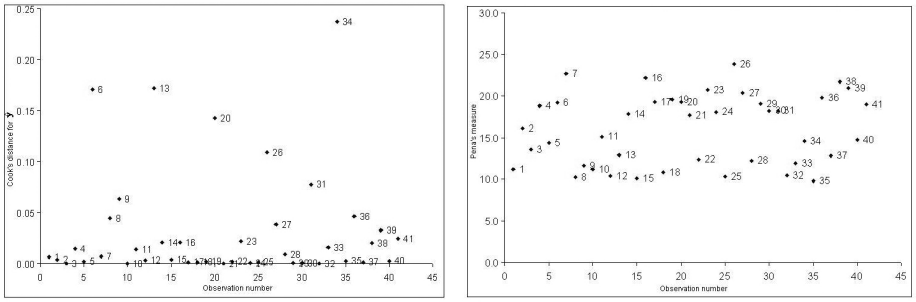


(a) (b)

FIGURE 1: (a) index plot of residuals, $\breve{e}_i$ (b) index plot of leverages values, $\breve{h}_{ii}$



(a) (b)



(c) (d)

FIGURE 2: Plots for diabetes data: (a) index plot of Cook's distance for $\widehat{\boldsymbol{\beta}}$, $\widetilde{C}_i$ (b) index plot of Cook's distance for $\widehat{\mathbf{m}}$, C_i^*, (c) index plot of Cook's distance for, $\widehat{\mathbf{y}}$, $\breve{C}_i$ (d) index plot of Pena's measure $\widetilde{S}_i$.

Table 1: Estimates of parametric and nonparametric components

Estimates of Parametric Component		Estimates of Nonparametric Component	
0.008	0.111	4.950	4.450
-0.501	0.312	5.206	5.345
0.339	0.261	05.279	5.319
-0.055	0.329	5.282	5.168
-0.539	-0.327	4.563	5.343
-0.711	0.286	5.332	5.342
-0.280	0.330	5.341	5.253
0.298	-0.430	5.003	5.295
0.366	0.323	4.617	5.327
0.033	-0.573	4.912	5.297
-0.369	0.181	5.156	4.941
0.213	-0.063	4.950	4.912
-0.079	-0.477	4.435	4.852
0.256	0.251	5.316	5.089
0.309	0.319	5.156	5.338
-0.133	0.210	5.309	5.257
-0.249	-0.407	5.282	5.329
0.404	0.251	5.191	5.338
0.036	-0.159	5.298	5.212
0.307	-0.382	5.333	5.289
0.176		5.304	

From Figure 2, according to Cook's distances ($\widetilde{C}$, C_i^* and $\breve{C}_i$) adjusted by Kim et al. (2002), observations 6, 34, 31, 20 and 26 are considered the five most influential observations on $\widehat{\boldsymbol{\beta}}$, observations 22, 13, 23, 26, 20 are considered the five most influential observations on $\widehat{\mathbf{m}}$ and observations 34, 6, 20, 26, 13 are considered the five most influential observations on $\widehat{\mathbf{y}}$. As seen from Figure 1(a), 1(b), there are no high leverage outliers in the data. Therefore, according to our adjusted Pena's measure $\widetilde{S}_i$, which is not useful in situations there are the outliers with low leverage, no observation is considered influential.

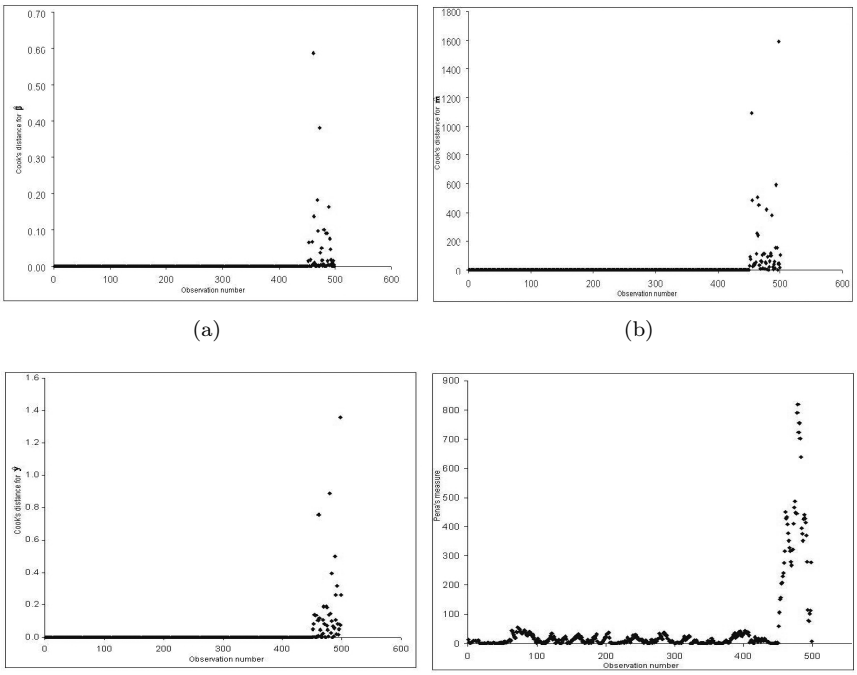
5.2. Artificial Data

Since we illustrate the performance of adjusted Pena's measure $\widetilde{S}_i$, an artificial data set with high leverage outliers is generated for semiparametric regression. We generate the data set using the model in the study of Kim et al. (2002)

$$y_i = 0.5\mathbf{z}_i + (x_i - 0.5)^2 + \varepsilon_i$$

We generate the 500 observations in which the last 50 observations would be high leverage outliers. For this reason, the first 450 of x_i from $U(0,1)$ and $z_i = i/450$ where ε_i is generated from $N(0, 0.02)$. The remaining 50 of x_i are generated from $U(5,10)$ and $z_i = i/50$ where ε_i is generated from $N(5,2)$. We suspect the last 50 observations for high leverage outliers. Figure 3 shows that the index plots of $\widetilde{C}$, C_i^*, $\breve{C}_i$ and $\widetilde{S}_i$.

As seen from Figure 3, $\widetilde{S}_i$ perfectly identifies 50 observations (observations $451 - 500$) as high leverage outliers. It is said that $\widetilde{S}_i$ is very useful for identifying high leverage outliers in semiparametric regression as in linear regression. In addition, $\widetilde{S}_i$ is clearly better than Cook's distances $(\widetilde{C}_i, C_i^*, \breve{C}_i)$ to detect high leverage outliers in large data as mentioned before.



(a) (b)

(c) (d)

FIGURE 3: Plots for Diabetes data: (a) index plot of Cook's distance for $\widehat{\boldsymbol{\beta}}$, $\widetilde{C}_i$ (b) index plot of Cook's distance for $\widehat{\mathbf{m}}$, C_i^*, (c) index plot of Cook's distance for, $\widehat{\mathbf{y}}$, $\breve{C}_i$ (d) index plot of Pena's measure $\widetilde{S}_i$.

5.3. Simulation Results

Here, we present a Monte Carlo simulation study that is designed to compare the performance of adjusted Pena's measure for semiparametric regression model. We generate the data sets from the same model in the previous section. We consider three different sample sizes, $n = 50$, 100, 250 with two different levels of influential observations ($i.e, \gamma = 10\%, 20\%$). The comparison of influence measures $(\widetilde{C}, C_i^*, \breve{C}_i$ and $\widetilde{S}_i)$ in semiparametric regression is carried out by the following steps:

1. Generation of the data with certain percentage of high leverages (X's outliers): For this purpose, we generate the first $n(1-\gamma)\%$ of x_i from $U(0,1)$

and $z_i = i/(n(1-\gamma)\%)$ where ε_i is generated from $N(0, 0.02)$. The remaining $n\gamma\%$ of x_i are generated from $U(5, 10)$ and $z_i = i/(n\gamma\%)$ where ε_i is generated from $N(0, 0.02)$.

2. Generation of the data with certain percentage of both high leverages (X's outliers) and outliers (Y's outliers): For this purpose, we generate the first $n(1-\gamma)\%$ of x_i from $U(0, 1)$ and $z_i = i/(n(1-\gamma)\%)$ where ε_i is generated from $N(0, 0.02)$. The remaining $n\gamma\%$ of x_i are generated from $U(5, 10)$ and $z_i = i/(n\gamma\%)$ where ε_i is generated from $N(5, 2)$.

3. Generation of the data with certain percentage of both intermediate-leverages and outliers (Y's outliers): For this purpose, we generate the first $n(1-\gamma)\%$ of x_i from $U(0, 1)$ and $z_i = i/(n(1-\gamma)\%)$ where ε_i is generated from $N(0, 0.02)$. The remaining $n\gamma\%$ of x_i are generated from $U(1, 3)$ and $z_i = i/(n\gamma\%)$ where ε_i is generated from $N(5, 2)$.

4. Generation of the data with certain percentage of low outliers: For this purpose, we generate the first $n(1-\gamma)\%$ of x_i from $U(0, 1)$ and $z_i = i/(n(1-\gamma)\%)$ where ε_i is generated from $N(0, 0.02)$. The remaining $n\gamma\%$ of x_i are generated from $U(1, 3)$ and $z_i = i/(n\gamma\%)$ where ε_i is generated from $N(1, 0.2)$.

5. Each measure is computed from each of the 100 replications.

6. Make comparison of detection of influential observations by using correct determination rate of each measure (i.e., total number of influential observations identified divided by total number of influential observations).

Table 2-5 show the correct determination rate of each measure ($\widetilde{C}, C_i^*, \breve{C}_i$ and $\widetilde{S}_i$) for different shows sizes and percentages of influential observations from 100 replications. From Table 2, adjusted Pena's measure, $\widetilde{S}_i$, performs similar results with Cook's distance $\breve{C}_i$ for $\widehat{y}$ to identify the high leverages for all the sample size. But, it is better than $\widetilde{C}_i$, C_i^* for all situations. From Table 3, adjusted Pena's measure, $\widetilde{S}_i$ clearly performs better than Cook's distances for $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{m}}$ and $\widehat{\mathbf{y}}$ ($\widetilde{C}_i$, C_i^*, $\breve{C}_i$) to detect high leverages outliers in large data. As seen from Table 3, almost all high leverage outliers could correctly be detected by $\widetilde{S}_i$ for $n = 250$. From Table 4, adjusted Pena's measure $\widetilde{S}_i$ successfully identifies intermediate leverage outliers that are not detected by Cook's distance for $n = 100$ and $n = 250$. From Table 5, adjusted Pena's measure $\widetilde{S}_i$ fails to detect low outliers with no high leverage as expected.

Table 2: The correct determination rate of high leverages (X's outliers).

Sample Size	Percentages of influential observations	Correct determination of measures (in percentages)			
		$\widetilde{C}_i$	C_i^*	$\breve{C}_i$	$\widetilde{S}_i$
n=50	10%	33	60	60	68
	20%	16	19	39	36
n=100	10%	23	11	39	45
	20%	17	14	38	35
n=250	10%	49	50	69	72
	20%	43	17	75	76

$\widetilde{C}_i$: Cook's distance for $\widehat{\boldsymbol{\beta}}$; C_i^*: Cook's distance for $\widehat{\mathbf{m}}$; $\breve{C}_i$: Cook's distance for $\widehat{\mathbf{y}}$; $\widetilde{S}_i$: Adjusted Pena's measure

Table 3: The correct determination rate of both high leverages (X's outliers) and outliers (Y's outliers).

Sample size	Percentages of influential observations	Correct determination of measures (in percentages)			
		$\widetilde{C}_i$	C_i^*	$\breve{C}_i$	$\widetilde{S}_i$
n=50	10%	51	70	72	80
	20%	46	44	68	84
n=100	10%	49	66	75	91
	20%	45	23	65	92
n=250	10%	52	52	71	98
	20%	44	19	62	98

$\widetilde{C}_i$: Cook's distance for $\widehat{\boldsymbol{\beta}}$; C_i^*: Cook's distance for $\widehat{\mathbf{m}}$; $\breve{C}_i$: Cook's distance for $\widehat{\mathbf{y}}$; $\widetilde{S}_i$: Adjusted Pena's measure.

Table 4: The correct determination rate of both intermediate leverages (X's outliers) and outliers (Y's outliers).

Sample size	Percentages of influential observations	Correct determination of measures (in percentages)			
		$\widetilde{C}_i$	C_i^*	$\breve{C}_i$	$\widetilde{S}_i$
n=50	10%	40	48	81	82
	20%	32	34	70	86
n=100	10%	32	39	77	86
	20%	23	27	66	89
n=250	10%	20	31	73	94
	20%	14	17	63	96

$\widetilde{C}_i$: Cook's distance for $\widehat{\boldsymbol{\beta}}$; C_i^*: Cook's distance for $\widehat{\mathbf{m}}$; $\breve{C}_i$: Cook's distance for $\widehat{\mathbf{y}}$; $\widetilde{S}_i$: Adjusted Pena's measure.

Table 5: The Correct Determination Rate of low outliers.

Sample size	Percentages of influential observations	Correct determination of measures (in percentages)			
		$\widetilde{C}_i$	C_i^*	$\breve{C}_i$	$\widetilde{S}_i$
n=50	10%	51	38	51	21
	20%	28	18	33	22
n=100	10%	39	43	47	13
	20%	25	19	30	4
n=250	10%	33	29	43	13
	20%	23	12	31	1

$\widetilde{C}_i$: Cook's distance for $\widehat{\boldsymbol{\beta}}$; C_i^*: Cook's distance for $\widehat{\mathbf{m}}$; $\breve{C}_i$: Cook's distance for $\widehat{\mathbf{y}}$; $\widetilde{S}_i$: Adjusted Pena's measure.

6. Conclusions

In this paper, we derived Pena's measure formula for semiparametric regression. The numerical examples and simulation study show that the proposed Pena's measure $\widetilde{S}_i$ performs very effectively in the identification of high leverage outliers and intermediate-leverage outliers in large data sets that are not clearly detected by adjusted Cook's distances for semiparametric regression model.

[Recibido: marzo de 2013 — Aceptado: junio de 2013]

References

Cook, R. (1977), 'Detection of influential observations in linear regression', *Technometrics* **19**, 15–18.

Kim, C. (1996), 'Cook's distance in spline smoothing', *Statistics and Probability Letters* **31**, 139–144.

Kim, C. & Kim, W. (1998), 'Some diagnostics results in nonparametric density estimation', *Communications in Statistics - Theory and Methods* **27**, 291–303.

Kim, C., Park, B. & Kim, W. (2001), 'Cook's distance in local polynomial regression', *Statistics & Probability Letters* **54**, 33–40.

Kim, C., Park, B. & Kim, W. (2002), 'Influential diagnostics in semiparametric regression models', *Statistics & Probability Letters* **60**, 49–58.

Pena, D. (2005), 'A new statistic for influence in linear regression', *Technometrics* **47**, 1–12.

Speckman, P. (1988), 'Kernel smoothing in partial linear models', *Journal of the Royal Statistical Society. Series B* **50**(3), 413–436.

Thomas, W. (1991), 'Influence diagnostics for the cross-validated smoothing parameter in spline smoothing', *Journal of the American Statistical Association* **86**(415), 693–698.

Türkan, S. (2012), Analysis of influential observation in semiparametric regression model, Doctoral Thesis, Hacettepe University, Faculty of Science. Department of Statistics, Ankara.

Türkan, S. and Toktamis, Ö. (2012), 'Detection of influential observations in ridge regression and modified ridge regression', *Model Assisted Statistics and Applications* **7**, 91–97.

Zhang, C., Mei, C. & Zhang, J. (2007), 'Influence diagnostics in partially varying-coefficient models', *Acta Mathematicae Applicatae Sinica* **23**(4), 619–628.

Zhu, Z. & Wei, B. (2001), 'Influence analysis in semiparametric nonlinear regression models', *Acta Mathematicae Applicatae Sinica* **24**(4), 568–581.