DOI: http://dx.doi.org/10.18185/eufbed.58091

DIAGNOSTICS OF INFLUENTIAL OBSERVATIONS IN PARTIALLY LINEAR MODELS

SEMİPARAMETRİK REGRESYON MODELLERİNDE ETKİLİ GÖZLEMLERİN TESPİTİ

Dursun AYDIN[1], Özlem GÜRÜNLÜ ALMA[2]

[1,2] Department of Statistics, Faculty of Sciences, Muğla Sıtkı Koçman University, Muğla,Turkey

Geliş Tarihi (Received): 02/01/2015 **Kabul Tarihi (Accepted)**: 25/03/2015

ABSTRACT

This paper considers the role of influence diagnostics in the partially linear regression models, $\mathbf{y} = \mathbf{X\beta} + \mathbf{f} + \boldsymbol{\varepsilon}$. An influential observation on the estimator of the coefficient vector $\boldsymbol{\beta}$ may not be influential on that of the nonparametric component f(x), and vice versa. Also, an observation which is not influential on either parametric or non-parametric component may be influential on the estimator of the mean response. So, we focus on influence measures for each estimator $\boldsymbol{\beta}$, $\mathbf{f}$, and the mean response $\mathbf{X\beta} + \mathbf{f}$. In the literature, the Cook's distance is used to detect influential observation in partially linear models. In certain types of data sets, it is quite common an unusual observation or a small subset using Dffits, Dfbetas, and CovRatio statistics. Therefore, in our study, Dffits, Dfbetas, and CovRatio are proposed to identify any influential observation in the partially linear regression models. These measures are discussed on each of which measures the effect of detecting an influential observation by using real and simulation data sets.

Keywords: Diagnostics measures, influential observation, smoothing spline, smoothing parameter, partially linear model (PLM).

ÖZET

Bu çalışma, yarı parametrik regresyon modellerindeki $\mathbf{y} = \mathbf{X\beta} + \mathbf{f} + \boldsymbol{\varepsilon}$ etkili gözlemlerin rolünü ele almaktadır. $\boldsymbol{\beta}$ katsayı vektörünün kestirimi üzerindeki etkili bir gözlem, modelin parametrik olmayan bileşeni f(x) üzerinde etkili olmayabilir, veya bunu tam tersini de söyleyebiliriz. Aynı zamanda, parametrik veya parametrik olmayan bileşen üzerinde etkili olmayan bir gözlem ortalama yanıt tahmincisinde etkili olabilir. Bu yüzden, her bir kestirici $\boldsymbol{\beta}$, $\mathbf{f}$ ve yanıt değişkeni için $\mathbf{X\beta} + \mathbf{f}$ etki ölçümleri üzerinde odaklanılmıştır. Literatürde, yarı parametrik regresyon modellerindeki etkili gözlemlerin tespit edilmesinde Cook's uzaklık ölçüsü kullanılmaktadır. Veri kümelerinin belirli türlerinde Dffits, Dfbetas ve

***Sorumlu Yazar:** *ozlem.gurunlu@gmail.com**

CovRatio istatistiklerinin etkili gözlemlerin tespitinde kullanımı yaygındır. Bu nedenle, çalışmamızda yarı parametrik regresyon modellerindeki herhangi bir etkili gözlemin tespit edilebilmesi için Dffits, Dfbetas, and CovRatio istatistikleri önerilmiştir. Etkili gözlemlerin tespit edilmesinde her bir ölçümün etkisi gerçek ve benzetim çalışması veri kümeleri kullanılarak ele alınmıştır.

Anahtar kelimeler: Etkili gözlem tanı ölçümleri, Etkili gözlem, Düzleştirme eğrisi, Düzleştirme parametresi, yarı parametrik regresyon modeli.

1.INTRODUCTION

Suppose that responses $y_1,...,y_n$ are depended on multiple explanatory variables. In a multiple linear regression setting, it is assumed that dependence among response and explanatory variables is linear and the theory of the general linear model can be used to estimate the model. The response and explanatory values are connected by the following regression model:

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

where $\boldsymbol{x}_{\mathrm{i}} = \left(\mathrm{x}_{\mathrm{i1}}, \mathrm{x}_{\mathrm{i2}}, ...\mathrm{x}_{\mathrm{ik}}\right)'$ and $x_1, x_2, ..., x_n$ are known $\mathrm{k}-$ dimensional vector with $\mathrm{k} \le \mathrm{n}$, $\{t_i\}$ are non-stochastic knot points of an explanatory variable t , $\boldsymbol{\beta}$ is an unknown $\mathrm{k}-$dimensional vector of regression coefficients, $\mathrm{f} \in \mathrm{C}^2$ $[\mathrm{a}, \mathrm{b}]$ is an unknown smooth function, and $\{\varepsilon_{\mathrm{i}}\}$ are random errors assumed to be i.i.d. $\mathrm{N}(0, \sigma^2)$ distributed. Further, it is assumed that $\mathrm{t}_1 \le \mathrm{t}_2 \le, ..., \le \mathrm{t}_{\mathrm{n}}$, where n is the number of observations in the sample. The main goal is to estimate vector of regression coefficient $\boldsymbol{\beta}$ and unknown non-parametric smooth function f(t) from the data set $\{\mathrm{y}_{\mathrm{i}}, \mathrm{x}_{\mathrm{i}}, \mathrm{t}_{\mathrm{i}}\}$. Model (1) is called partially linear model. In vector-matrix form, the model can be expressed as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon} \qquad (2)$$

where

$\mathbf{y} = \left(\mathrm{y}_1, ...\mathrm{y}_{\mathrm{n}}\right)^{\mathrm{T}}$, $\mathbf{f} = \left(\mathrm{f}(\mathrm{t}_1), ..., \mathrm{f}(\mathrm{t}_{\mathrm{n}})\right)^{\mathrm{T}}$, $\boldsymbol{\varepsilon} = \left(\varepsilon_1, ...\varepsilon_{\mathrm{n}}\right)^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}} = \left(x_1, ..., x_n\right)$ is the $\mathrm{n} \times \mathrm{k}$ matrix. Partially linear models have been widely used in various applications. They allow easier interpretation of the effect of each variable and preferable to a completely nonparametric model since the well-known reason curse of dimensionality. In addition, these models are most useful devices for extracting and understanding the essential features of datasets. However, most of the databases in real-world include a particular amount of abnormal values, generally termed as outliers. An accurate identification of outliers plays a significant role in statistical analysis especially regression models. Nevertheless, many classical statistical models are blindly applied to data sets containing outliers; the results can be misleading at best. The appearance of outliers can exert negative influences on the fit of the multiple regression models.

The study of outliers has interested practicing statisticians and other scientists for a great number of years. Thompson (1935) was the first author to drop both assumptions about population mean and standard deviation. Anscombe (1960) and Daniel (1960) were among the first authors to propose the use of standardized residual for detecting a single outlier in linear regression models. Most of the regression diagnostics have been concerned about parametric regression models. In the classical linear models, an important approach for identifying influential observations based on case deletion was proposed by Cook (1977).Cook's distance measures the effect of removing one observation on a parameter estimate or a fitted value. Kim and Storer (1996) studied reference values for Cook's distance. Diagnostics studies for generalized linear models can be found in Thomas and Cook (1989), Davison and Tsai (1992). In nonparametric regression models, Eubank [1984, 1985], Silverman (1985), Carmody (1988), Thomas (1991) and Kim (1996) investigated influence measures for splines. Kim et.al., (2001) suggested a type of Cook's distance in local polynomial regression. Kim et.al., (2002) studied influence of observations on some estimators of the parametric and the nonparametric components in the semi-parametric model. They observe that the influence diagnostics in the semi-parametric model have different aspects from those in the parametric and the nonparametric models. They expressed different aspects of semi-parametric regression diagnostics such as $\tilde{C}_i, C_i^*$ and C_L . Fung et.al., (2002) considered both case and subject deletion diagnostics, as well as outlier screening for semi-parametric mixed models. They used Cook's distance to measure influence on a linear parameter estimate and DFFIT to measure changes in a nonparametric fit. In this paper, we are interested in the identification of influential points in partially linear models. For this purpose, we are generalized Dffits, Dfbetas, and CovRatio statistics for partially linear models. Then, we are compared performances of diagnostic statistics whether truly finding of influential observations.

The rest of the paper organized as follows: The estimation of semi-parametric regression models are defined in Section 2. Also Cook's distances, Dffits, Dfbetas, and CovRatio statistics are given in this section. Illustrative examples based on real data sets and Monte Carlo simulations are given in Section 3. Finally, we make our experimental evaluation and concluding remarks in Section 4.

2.METHODS

2.1. Estimators of Partially Linear Models

Let $\mathbf{f} = (f(t_1),...,f(t_n))$ be the vector of values of function f at the knot points $t_1,...,t_n$. The smoothing spline estimate $\hat{\mathbf{f}}_\lambda$ of this vector or the fitted values for data $y = (y_1,...,y_n)^T$ are projected by

$$\hat{\mathbf{f}} = \left(\hat{f}_\lambda(t_1),...,\hat{f}_\lambda(t_n)\right)^T = \left(S_\lambda\right)\left(y_1,...,y_n\right)^T \text{ or, in short, } \hat{\mathbf{f}}_\lambda = S_\lambda \mathbf{y}. \quad (3)$$

where $\hat{f}_\lambda$ is a natural cubic spline with knots at $t_1,...,t_n$ for a fixed smoothing parameter $\lambda > 0$, and S_λ is a well-known positive-definite (symmetrical) smoother matrix which depends on λ and the knot points $t_1,...,t_n$, but not on $\mathbf{y}$. Function $\hat{f}_\lambda$, the estimator of function f, is obtained by cubic spline interpolation that rests on condition $\hat{f}(t_i) = (\hat{\mathbf{f}})_i$, $i = 1,2,...,n$. To gain better perspective on smoothing spline, Eubank (1999), Green and Silverman (1994), Wahba (1990) state studied opinions.

For smoothing spline based estimation of the parameters of interest in PLM a solution can be performed by minimizing the following sum of squares equation,

$$SS(\boldsymbol{\beta}, f) = \sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - f(t_i)\right)^2 + \lambda\int_a^b \left(f''(t)^2\right)dx \quad (4)$$

where $f \in C^2[a,b]$. $\boldsymbol{\beta} \in R^k$ and $\mathbf{x}_i$ is the ith row of the matrix $\mathbf{X}$. The resulting estimator is called as smoothing spline. On the other hand, Equation (4) is also known as the roughness penalty approach Green and Silverman (1994). This estimation concept is based on iterative solution of the normal equations. They propose to use backfitting algorithm. Below, we present an alternative to second concept of Green and Silverman (1994), a direct method. Rice (1986) indicated that partial spline estimator is generally biased for the optimal λ choice when the components of $\mathbf{X}$ on t. This asymptotic bias can be larger than the standard error. For notational convenience, let X be the matrix with x_i' as the ith row, and y be a response vector. Let $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}$ and $\tilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Y}$. Then $\mathbf{r} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ is the residual vector from the linear terms. Thus, matrix $\tilde{\mathbf{H}} = \mathbf{H} - \mathbf{H}^*$, $\mathbf{H}^* = \mathbf{S}_\lambda(\mathbf{I} - \mathbf{H})$ can be defined as hat matrix. Then we can write $\mathbf{X}\hat{\boldsymbol{\beta}} = \tilde{\mathbf{H}}\mathbf{Y}$ and $\hat{\mathbf{f}}(\mathbf{t}) = \mathbf{S}_\lambda(\mathbf{I}\text{-}\tilde{\mathbf{H}})\mathbf{Y}$, respectively. Also the vector of the fitted responses equals $\hat{\mathbf{Y}} = \mu = \mathbf{Hy},$ and is expressed as follows:

$$
\begin{aligned}
\mu_s &= \hat{f}_s + \mathbf{X}\hat{\boldsymbol{\beta}}_s, \\
&= \mathbf{S}_\lambda(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_s) + \mathbf{X}\hat{\boldsymbol{\beta}}_s \\
&= \mathbf{S}_\lambda\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_s\mathbf{S}_\lambda + \mathbf{X}\hat{\boldsymbol{\beta}}_s \\
&= \mathbf{S}_\lambda\mathbf{Y} + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}\hat{\boldsymbol{\beta}}_s \\
&= \mathbf{S}_\lambda\mathbf{Y} + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}(\mathbf{I} - S_\lambda)\mathbf{Y} \\
&= \mathbf{Y}(S_\lambda + \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}(\mathbf{I} - \mathbf{S}_\lambda)) \\
&= \mathbf{S}_\lambda + \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Y} \\
&= \mathbf{HY},
\end{aligned}
$$

where $\mathbf{H} = \mathbf{S}_\lambda + (\mathbf{I} - \mathbf{S}_\lambda)\tilde{\mathbf{H}} = \mathbf{S}_\lambda + \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}(\mathbf{I}\text{-}\mathbf{S}_\lambda)$ is the hat matrix. Note that $\mathbf{H} = \tilde{\mathbf{H}} + \mathbf{H}^* = \mathbf{S}_\lambda + \tilde{X}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{S}_\lambda)$ which will be used in defining and interpreting Cook's distances in semi-parametric model. On the other hand, hat matrices for parametric and nonparametric components of semiparametric model (1) can be defined as following way;

$$
\begin{aligned}
\mu_s &= \mathbf{X}\hat{\boldsymbol{\beta}}_s + \hat{f} \\
&= X(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{Y}} + \mathbf{S}_\lambda(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_s) \\
&= X(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}(\mathbf{I}\text{-}S_\lambda)\mathbf{Y} + S_\lambda(Y - X(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{Y}}) \\
&= \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}\mathbf{Y}+S_\lambda(Y - X(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}(\mathbf{I}\text{-}\mathbf{S}_\lambda)\mathbf{Y}) \\
&= \tilde{\mathbf{X}}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}\mathbf{Y}+S_\lambda(Y - \tilde{X}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}\mathbf{Y}) \\
&= \tilde{\mathbf{H}}\mathbf{Y}+\mathbf{S}_\lambda(\mathbf{I}\text{-}\tilde{\mathbf{H}})\mathbf{Y} \\
\mu_s &= \mathbf{HY}=\mathbf{Y}(\tilde{\mathbf{H}}+\mathbf{S}_\lambda(\mathbf{I}\text{-}\tilde{\mathbf{H}}))
\end{aligned}
$$

where $\mathbf{H} = \tilde{\mathbf{H}} + \mathbf{S}_\lambda(\mathbf{I} - \tilde{\mathbf{H}})$ or $\mathbf{H} = \tilde{\mathbf{H}} + \mathbf{H}^*$ can be defined as hat matrix for model (1). Thus, $\mathbf{H}^* = \mathbf{S}_\lambda(\mathbf{I} - \tilde{\mathbf{H}})$ and $\tilde{\mathbf{H}} = \mathbf{H} - \mathbf{H}^*$ are defined as hat matrices for nonparametric and parametric components of model (1), respectively.

In this paper, we considered the Speckman's smoothing spline approach given by four steps. Applying results due to Speckman (1988) this bias can be substantially reduced. Respective estimating process as follows:

Step 1: Given a smoother matrix S_λ, depending on smoothing parameter λ construct the residuals $\tilde{\mathbf{X}} = (I - S_\lambda)\mathbf{X}$ and $\tilde{\mathbf{Y}} = (I\text{-}S_\lambda)\mathbf{Y}$, respectively.

Step 2: For parametric component of the equation (1) the vector of the regression coefficients $\boldsymbol{\beta}$ can be estimated by regressing the residuals of $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}$. So, the $\hat{\boldsymbol{\beta}}$ is given by

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

Step 3: By substitution of the equations (5) in (1), construct $\mathbf{y}_i^* = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$. In this case, the PLM in equation (1) is obtained as $\mathbf{y}_i^* = f(t_i) + \varepsilon_i$ then, obtain an estimate $\hat{\mathbf{f}}$ by smoothing spline method applied to values $\mathbf{y}_i^*$. For the nonparametric component of the equation (1) the vector of the $\hat{\mathbf{f}}$ is obtained as follows:

$$\hat{\mathbf{f}} = S_\lambda(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{6}$$

Step 4: Evaluate some selection criteria functions such as Akaike information criterion (AIC), cross validation (CV), Mallow's

Cp, generalized cross validation (GCV) and risk estimation classical plots (REC) and iterate, changing λ , until it is minimized. In our study, we used GCV,

$$\mathrm{GCV}=\sum_{i=1}^{n} e_i^2 \Big/ \left\{1-\frac{1}{n}\operatorname{tr}(\mathbf{H})\right\}^2 .$$

As can be seen, the estimates are obtained by regression on partial residuals. In Speckman (1988) approach estimator of $\boldsymbol{\beta}$ is obtained after removing the effect of t from both the $\mathbf{x}_i$ and y .

2.2. Detecting Influential Observations in Partially Linear Models

In this section, we consider the influence of a single observation on the estimator $\hat{\boldsymbol{\beta}}, \hat{\mathbf{f}},$ and $\hat{\mathbf{y}}$. An influence measure for the i^{th} observation on $\hat{\boldsymbol{\beta}}, \hat{\mathbf{f}},$ and $\hat{\mathbf{y}}$ were defined as a type of Cook's distances by Kim et. al., (2002). Cook's distance for influence on $\hat{\boldsymbol{\beta}}$ by

$$\tilde{C}_i = (\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathbf{i}})'\tilde{\mathbf{X}}'\tilde{\mathbf{X}}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{\mathbf{i}})/\sigma^2\operatorname{tr}(\tilde{\mathbf{H}}). \tag{7}$$

Nothing $\operatorname{tr}(\tilde{\mathbf{H}})=\mathrm{k}$, it can be expressed as a function of the i^{th} residual and leverage,

$$\tilde{C}_i = \frac{1}{k\sigma^2}\frac{\tilde{e}_i\tilde{h}_{ii}}{(1-\tilde{h}_{ii})^2} \tag{8}$$

where $\tilde{e}_i$ is the i^{th} component of residual vector $\tilde{\mathbf{e}}=(\mathbf{I}-\tilde{\mathbf{H}})\tilde{\mathbf{y}}$ and $\tilde{h}_{ii}$ is the i^{th} diagonal component of $\tilde{\mathbf{H}}$, and an estimator of σ^2 is $s^2=\sum_{i=1}^{n} e_i^2/\{n-\operatorname{tr}(\mathbf{H})\}$. The Cook's distance for $\hat{\mathbf{f}}$ according to the types of local polynomial smoother for the i^{th} observation by

$$\begin{gathered} C_i^* = \left\{\hat{f}(t_i)-\hat{f}_i(t_i)\right\}^2/\sigma^2\operatorname{tr}(\mathbf{H}^*), \\ C_i^* = (h_{ii}^* e_i^*)^2/\left\{(1-h_{ii}^2)\sigma^2\operatorname{tr}(\mathbf{H}^*)\right\}, \end{gathered} \tag{9}$$

where h_{ii}^* is the i^{th} diagonal element of $\mathbf{H}^*$ and e_i^* is the i^{th} component of residual vector $\mathbf{e}^*=(\mathbf{I}-\mathbf{H}^*)\mathbf{y}$. An influence measure for the i^{th} observation on the vector of fitted values can be similarly defined by

$$C_i = (\hat{\mathbf{y}}-\hat{\mathbf{y}}_i)'(\hat{\mathbf{y}}-\hat{\mathbf{y}}_i)/\sigma^2\operatorname{tr}(\mathbf{H}). \tag{10}$$

It can be expressed as a function of the corresponding residual and leverage. Let h_{ii} is the i^{th} diagonal element of $\mathbf{H}$ and e is the i^{th} component of residual vector $\mathbf{e} = (\mathbf{I}-\mathbf{H})\mathbf{y}$. The influence measure fort the l observations with indices in the set $L=\{i_1,\ldots,i_l\}$ may be defined by

$$\begin{aligned} C_L &= (\hat{\mathbf{y}}-\hat{\mathbf{y}}_L)'(\hat{\mathbf{y}}-\hat{\mathbf{y}}_L)/\sigma^2\mathrm{tr}(\mathbf{H}) \\ &= \mathbf{e}_L'(\mathbf{I}-\mathbf{H}_L)^{-1}\mathbf{H}_L(\mathbf{I}-\mathbf{H}_L)^{-1}\mathbf{e}_L/\sigma^2\mathrm{tr}(\mathbf{H}), \end{aligned} \tag{11}$$

where $e_L = (e_{i1},\ldots,e_{il})'$ and $\mathbf{H}_L$ is $l\times l$ submatrix of $\mathbf{H}$. Cook's distance measures the influence of the i^{th} observation if it is removed from the sample. Two other measures similar to Cook's distance are introduced by Belsley, Kuh, and Welch (1980) for linear regression models. We have also generalized these measures to partially linear model (1). The first of these is a statistics that indicates how much an observation has affected its fitted value from the semi-parametric regression model. In this study we are generalized Dffits, Dfbetas and CovRatio for semi-parametric regression models. Firstly, Dffits generalized as follows,

$$\mathrm{Dffits}_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}\frac{e_i}{(S_{(i)^2}(1-h_{ii}))^{1/2}}, \qquad i=1,2,\ldots,n \tag{12}$$

$$S_{(i)}^2 = \frac{(n-k)\hat{\sigma}^2-\left(\dfrac{e_i^2}{1-\tilde{h}_{ii}}\right)}{n-k-1} \tag{13}$$

where h_{ii} is the diagonal element of hat matrix, $\mathbf{H}=\mathbf{S}_\lambda+(\mathbf{I}-\mathbf{S}_\lambda)\widetilde{\mathbf{H}}$, k is the number of explanatory variables, and $S_{(i)}^2$ is estimate of σ^2 based on a data set with i^{th} observation removed. Thus, Dffits_i is the number of standard deviations that the fitted value changes if observation i is removed. A larger value of Dffits_i implies that the i^{th} observation may be a potential outlier. That is, if $|\mathrm{Dffits}_i| > 2\sqrt{\mathrm{tr(H)}/n}$, then the i^{th} observation is considered highly influential point. Note that Dffits_i is the number of standard errors that the fitted value $\hat{y}_i$ changes if case i is removed. The second of these is a statistics that indicates how much the regression coefficient $\hat{\beta}_j$ changes, in standard deviation units, if the i^{th} case were deleted. The Dfbetas statistics may be written as

$$\text{Dfbetas}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{s^2_{(i)}(\tilde{X}'\tilde{X})^{-1}}} \tag{14}$$

$\text{Dfbetas}_{j(i)}$ implies that the i^{th} observation may be a potential outlier and it tells the change of regression parameter estimate $\hat{\beta}_j$ due to the deletion of the i^{th} observation. As suggested by Belsley, Kuh, and Welch (1980), the cutoff values are $2/\sqrt{n}$ for $\text{Dfbetas}_{j(i)}$ to detect possible outliers. If $|\text{Dfbetas}_{j(i)}| > 2/\sqrt{n}$, i^{th} observation is considered potential outliers.

Cook's distance, Dffits_i and $\text{Dfbetas}_{j(i)}$ provide a concept about the effect of observations on the fitted values and estimated regression coefficient. They do not provide any information about overall precision of estimation. To state the role of the i^{th} observation precision of estimation, CovRatio statistics can be used in applications. CovRatio provides a scalar measure of the impact of each observation on the variances (or standard errors) of the regression coefficients and their covariances. It is defined as,

$$\text{CovRatio}_i = \left(\frac{S_i^2}{\hat{\sigma}^2}\right)^{tr(\tilde{H})} \frac{1}{(1-\tilde{h}_{ii})} \tag{15}$$

Belsley, Kuh and Welsch (1980) proposed a cutoff value for the CovRatio . Note that if $\text{CovRatio}_i > 1+3tr(H)/n$ or if $\text{CovRatio}_i < 1+3tr(H)/n$ then i^{th} observation should be considered influential. On the other words, If there are values outside interval $[1 \pm 3tr(H)/n]$, then i^{th} observation may be identified as a possible outlier.

3.FINDINGS

In this section, we used two real data examples and a simulation experiment to illustrate the effectiveness of diagnostic measures.

3.1. Real data application 1: Diabetes Data

We used Diabetes data from a study by Sockett et. al., (1987). The response is the logarithm of C-peptide concentration at diagnosis and two predictors are x_1=age and x_2=base deficit. The semiparametric model is $y = 0.001 + 0.0554x_2 + f(x_1) + \varepsilon$. The local linear smoother is used and the bandwidth $\lambda = 5.6$ was selected by minimizing the GCV. The cut off formulas and their values concerning the statistics are presented in Table 1. Influential observations are identified by considering these threshold values for Diabetes data.

Table 1. The cut off formulas and their values of influence statistic

Influence Statistics	The cut off formulas	The cut off value
Cook's distances	$4/n$	0.097
Dffits	$2\sqrt{\mathrm{tr}(\mathbf{H})/n}$	0.785
Dfbetas	$2/\sqrt{n}$	0.312
CovRatio	$1 \pm 3\mathrm{tr}(\mathbf{H})/n$	< 0.537 or, > 1.462

The resulting values of Cook's distances, Dffits, Dfbetas, and CovRatio for influential observation are listed in Table 2. It illustrates the ability of the diagnostic statistics in detecting influential observations for all situations.

Table 2. Resulting values of influential observation statistics for Diabetes data

Obs. Number	$\tilde{C}$	C^*	C_L	Dffits	Dfbetas	CovRatio
1	0.001	0.000	0.001	0.103	0.004	1.461
2	0.028	0.008	0.013	-0.425	0.163	**1.472**
3	0.001	0.001	0.001	-0.105	0.034	1.443
4	0.036	0.018	0.016	0.544	0.008	1.173
5	0.020	0.002	0.006	0.251	0.068	**1.711**
6	**0.316**	0.050	**0.133**	**-1.398**	**0.630**	1.060
7	0.081	0.012	0.042	0.546	0.080	**1.949**
8	0.046	0.030	0.029	0.674	0.105	1.268
9	0.055	0.026	0.036	-0.643	0.158	**1.510**
10	0.007	0.002	0.003	-0.206	0.001	1.450
11	0.018	0.009	0.011	0.393	0.124	1.459
12	0.000	0.000	0.000	-0.012	0.002	**1.476**
13	**0.124**	0.020	0.069	-0.765	0.006	**1.757**
14	0.050	0.020	0.023	-0.645	0.169	1.175
15	0.030	0.014	0.013	-0.459	0.082	1.358

16	0.045	0.020	0.020	0.609	0.056	1.119
17	0.005	0.003	0.003	-0.218	0.045	1.375
18	0.000	0.000	0.000	0.028	0.008	**1.481**
19	0.008	0.004	0.004	-0.278	0.024	1.328
20	**0.322**	**0.129**	**0.150**	**1.885**	**0.407**	**0.494**
21	0.000	0.000	0.000	-0.045	0.010	1.408
22	**0.110**	0.009	0.039	0.598	0.050	**1.784**
23	0.077	0.024	0.038	0.727	0.125	1.369
24	0.001	0.001	0.001	-0.115	0.031	1.418
25	0.000	0.000	0.000	0.002	0.000	**1.466**
26	**0.207**	0.093	**0.100**	**-1.293**	0.303	0.968
27	0.094	0.013	0.043	-0.601	0.066	**1.817**
28	0.032	0.019	0.017	-0.528	0.146	1.303
29	0.003	0.001	0.002	-0.171	0.059	1.428
30	0.000	0.000	0.000	-0.061	0.018	1.445
31	**0.120**	0.016	0.046	**0.872**	**0.358**	1.164
32	0.018	0.004	0.006	-0.314	0.033	1.421
33	**0.101**	0.032	0.044	**-0.853**	0.058	1.152
34	**0.394**	**0.160**	**0.198**	**1.784**	**0.478**	0.829
35	0.000	0.000	0.000	0.003	0.000	1.453
36	**0.110**	0.050	0.054	**0.968**	**0.327**	1.062
37	0.007	0.005	0.004	-0.253	0.050	1.381
38	0.038	0.013	0.017	-0.524	0.169	1.322
39	0.055	0.027	0.026	-0.659	0.127	1.226
40	0.004	0.002	0.002	0.181	0.029	1.418
41	0.041	0.010	0.016	0.528	0.161	1.263

In Table 2, influential observations in Diabetes data are shown by bold characters. As can be seen from the Table 2 $\tilde{C}, C_L$, Dffits and Dfbetas methods are mostly detect the same observations as influential observations. Numbers of these observations are 6, 20, 26, 31, 34 and 36. C^* is only detected 20 and 34 observations on **y**, so it is not as effective as other diagnostics. CovRatio statistics is detected ten observations as influential observations which are effect the fitted values and estimated regression coefficient. These observations are 2, 5, 7, 9, 12, 13, 18, 22, 25, and 27.

3.2. Real data application 2: Windsor Data

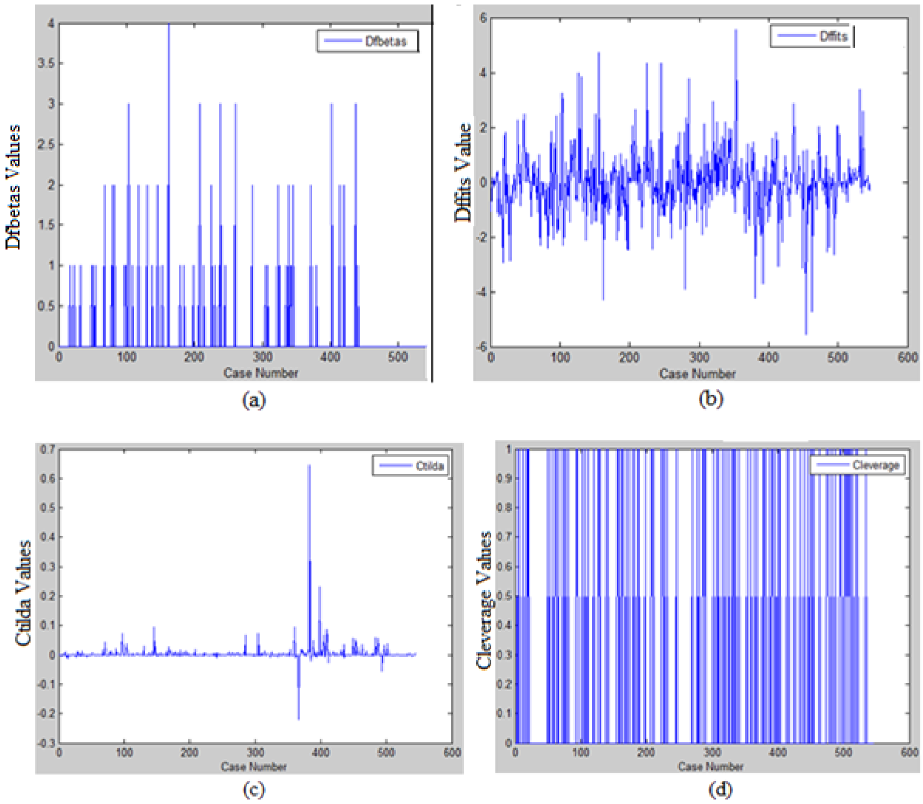
Anglin and Gencay (1996) data set was used as a second data set. It describes residential houses sold during July, August, and September of 1987 through the local Multiple Listing Service in Windsor which is a Canadian city. The 546 records contain information describing the key features of each house. The variables are defined as follows.

• DRV=1, if the house has a driveway.	• REC=1, if the house has a recreational room
• FFIN=1, if the house has full and finished basement.	• GHW=1, if the house uses gas for hot water heaing.
• CA=1, if there is a central conditioning	• GAR, shows the number of garage places
• REG=1, if the house located. Riverside or South Windsor	• LOT, lot size of the property in square feet.
• BDMS, the number of bedrooms.	• FB, the number of full bathrooms.
• STY, the number of stories.	

The specification of the semiparametric regression model is:

$$\begin{aligned}\log(P_i) &= 0.1313DRV_i + 0,0732REC_i + 0,0787FFIN_i + 0,2261GHW_i + 0,2244CA_i + 0,0809GAR_i \\ &+ 0,1434REG_i + f(LOT_i, BDMS_i, FB_i, STY_i) + \varepsilon_i \end{aligned} \tag{16}$$

where the mean of ε_i conditional on the explanatory variables is zero. The local linear smoother is used and we choose the bandwidth parameter by GCV as described in Section 2.1. The bandwidth parameter is minimized at $\lambda = 8$ which we used in the estimation of the semiparametric regression model in equation (16). Influence diagnostics statistics that was found by the number of observations for these data set, Dfbetas=79, Dffits=100, CovRatio=174, $\tilde{C} = 75$, $C^* = 9$, $C_L = 117$ respectively, when C^* did not find any influential observation. However, Dfbetas, Dffits. $\tilde{C}$ and C_L found the same 14 observations as influence observations. Their case numbers are 104, 131, 147, 157, 163, 209, 226, 261, 286, 354, 403, 415, 455, 464 and Figure 1 (a)-(d) show the results of Dfbetas, Dffits. $\tilde{C}$ and C_L statistics according to case number for Windsor data.



(a) (b) (c) (d)

Figure 1. (a)-(d): Results of Influence diagnostic statistics values according to case number for Windsor data

3.2. The simulation experiment

A Monte Carlo simulation study based on the above generalized diagnostics methods for semi-parametric regression models were carried out. The main goal was to study the different sized sample behavior of the generalized diagnostics methods for parametric and nonparametric components in semi-parametric models with different dimension. For the simulations we considered three examples of the semiparametric regression model which are defined as Equation (1) these are:

$$y_i = x'_{i1}\beta_1 + f(t_i) + \varepsilon_i,\ i=1,2,\ldots,n \tag{17}$$

$$y_i = x'_{i1}\beta_1 + x'_{i2}\beta_2 + x_{i3}\beta_3 + f(t_i) + \varepsilon_i \tag{18}$$

$$y_i = x'_{i1}\beta_1 + x'_{i2}\beta_2 + x'_{i3}\beta_3 + x'_{i4}\beta_4 + x'_{i5}\beta_5 + f(t_i) + \varepsilon_i \qquad (19)$$

The simulations setup is given in Table 3 and the performance of the diagnostics methods are compared for each of models and sample sizes when number of iteration is one thousand for each design.

Table 3. Specification of the simulations

Semiparametric Model	Equation 17	Equation 18	Equation 19
Parametric Component	$x_1 \sim N(5,1)$, $\beta = [0.5]$	$x_{1,2,3} \sim N(1, 0.001)$, $\beta = [0.5,1,1.5]$	$x_{1,2,3,4,5} \sim N(1, 0.001)$, $\beta = [0.5,1, 0.5,1.51.5]$
Nonparametric Component	$f(t_i)$=0.5sin($2\pi t$), t=1,2,...,n	$f(t_i)$=0.5sin($2\pi t$), t=1,2,...,n	$f(t_i)$=0.5sin($2\pi t$), t=1,2,...,n
ε_i	$\varepsilon_i \sim N(1,1)$	$\varepsilon_i \sim N(1,1)$	$\varepsilon_i \sim N(1,1)$
Added Influential Observation	X_1 and Y	X_1, X_2, X_3 and Y	X_1, X_2, X_3, X_4, X_5 and Y
n	25, 50, 100, 200, 400		

The semiparametric regression models were generated using parametric and nonparametric components from Table 3. Then, to obtain leverage point, we changed response and explanatory variables values. For this aim the influential observation density level selected as 5%. Tablo 4 shows the number of influential observations must be added $\mathbf{X}_i$, and $\mathbf{Y}$.

Tablo 4. The number of influential observations (#io) must be added in $\mathbf{X}_i$, and $\mathbf{Y}$ for each sample sizes

	n				
	25	50	100	200	400
#io	2	3	5	10	20

Firstly, semiparametric models were generated for each sample sizes. After the generated semiparametric models, the influential observations were generated from uniform distributions which lay at least +3 standard deviations from the mean of variables $\mathbf{X}_i$, and $\mathbf{Y}$ taking into account of the percentage of influential observations. And then, response and explanatory observations were changed by influential observations.

4. RESULTS AND CONCLUSIONS

In this study, the performances of Dffits, Dfbetas and CovRatio statistics against Cook's distances are evaluated and these statistics to influential observation detection is demonstrated through the developed simulation experiments. Each data set contain a known percentage of influential observations, diagnostic statistics exceptionally detected these observations in all data sets tested. The simulation results are reported in Table 5 and the values are the percentage of influential observations for 1000 replicates. The True score shows the performance of diagnostics statistics truly find influential observations. The False score shows the performance of diagnostics statistics failure to find influential observations. These are calculated by (20) formulas.

$$\text{False}=\frac{\begin{array}{c}\text{Total number of incorrectly identified influential observations +}\\ \text{Total number of failure to identified influential observations}\end{array}}{\text{Total number of observations}}\text{x}100\% \tag{20}$$

$$\text{True}=\frac{\text{Total number of correctly identified influential observations}}{\text{Total number of observations}}\text{x}100\%$$

Tablo 5. Simulation Results based on performance ratio for Semi-parametric Regression Models and Sample Sizes.

n	Semiparametric Model		Dfbetas	Dffits	CovRatio	Ctilda	Cleverage
25	17	True	100	100	100	100	100
		False	0	0	0.2	0	0
	18	True	100	100	76	100	100
		False	0	0	0	3	3
	19	True	100	100	100	100	100
		False	0	0	0	0	0
50	17	True	100	100	100	100	100
		False	0	0	0.5	0	0
	18	True	100	100	100	100	100
		False	0	0	0.2	3	3
	19	True	100	100	100	100	100
		False	0	0	0	0	0
100	17	True	100	100	100	100	100
		False	0	0	0.8	0	0
	18	True	100	100	100	100	100
		False	0	0	0	0	0
	19	True	100	100	100	100	100
		False	0	0	0	0	0
200	17	True	100	100	100	100	100
		False	0	0	0.13	0	0
	18	True	100	100	100	100	100
		False	0	0	0	0	0
	19	True	100	100	100	100	100
		False	0	0	0	0	0
400	17	True	100	100	100	100	100
		False		0	0.17	0	0
	18	True	100	100	100	100	100
		False	0	0	0	0	0
	19	True	100	100	100	100	100
		False	0	0	0	0	0

In this study, we consider the semiparametric regresion model $y = \mathbf{X}'\boldsymbol{\beta} + f + \varepsilon$, and we generalized Dffits, Dfbetas, and CovRatio influence diagnostics for estimators of $\boldsymbol{\beta}$, f, and the mean response $\mathbf{X}'\boldsymbol{\beta} + f$. Generalized diagnostic methods on real data and simulation work can be said to be successful in finding influential observations. Especially in the small samples, compared to the others failed to CovRatio and C^* statistics. However, it had improved with increasing sample size. CovRatio and methods other than C^* shared the same success. It is said that, does not matter the dimension of the semiparametric regression model on the performance of influence diagnostics for detecting of the influential observations.

REFERENCES

Anglin P.M., Gencay R. 1996. Semiparametric estimation of a hedonic price function. Journal of Econometrics. 11, 633-648.

Anscombe. F. J.1960. Rejection of outliers. Technometrics. 2, 123-147.

Belsley D. A., Kuh E., Welsch R. E. 1980. Regression diagnostics: identifying influential data and source of collinearity. New York: John Wiley.

Carmody T.J. 1988. Diagnostics for multivariate smoothing splines. Journal of Statistical Planning and Inference. 19(2), 171-186

Cook. R.D. 1977. Detection of influential observations in linear regression. Technometrics. 19, 15-18.

Daniel. C. 1960. Locating outliers in factorial experiments. Technometrics. 2, 149-156.

Davison A. C.. and Tsai. C.L. 1992. Regression model diagnostics. Int. Statistical Review. 60, 337-353.

Eubank R. L. 1984. The hat matrix for smoothing splines. Statistics and Probability Letters. 2, 9-14.

Eubank R.L. 1985. Diagnostics for smoothing splines. Journal of Statistical Society B. 47, 332-341.

Eubank, R.L. 1999. Nonparametric Regression and Smoothing Spline, Marcel Dekker Inc., New York.

Fung W-K., Zhu Z-Y, Wei B-C, He X. 2002. Journal of statistical society B. 64(3), 565-579.

Green. P.J., B.W. Silverman. 1994. Nonparametric regression and Generalized Linear Models. New York: Chapman & Hall.

Kim C. 1996. Cook's distance in spline smoothing. Statistics and Probability Letters. 31, 139-144.

Kim C., Lee Y., Park B.U.2001. Cook's distance in local polynomial regression. Statistics Probability Letters. 54, 33-40.

Kim C., Park B.U., Kim W. 2002. Influence diagnostics in semiparametric regression models. Statistics Probability Letters. 60, 49-58.

Kim C., Storer B.E. 1996. Reference values for Cook's distance. Communication in Statistics Simulation and Computation. 25, 691-709.

Rice, J. 1986. Convergence rates for partially spline models, Statis. Prob. Lett. 4, 203-208

Silverman B. W. 1985. Some aspects of the spline smoothing approach to nonparametric regression curve fitting. Journal of Statistical Society B. 47, 1-52.

Sockett. E.B., Daneman. D., Clarson. C., Ehrich.1987. R.M. Factors affecting the patterns of residual insulin secretion during the first year of type (I) diabetes mellitus in children. Diabetes. 30, 453-459.

Speckman. P.E. 1988. Regression analysis for partially linear models. Journal Royal Statistics ser. B. 50, 413-436.

Thomas W. 1991. Influence diagnostics for the cross-validated smoothing parameter in spline smoothing. J. American Statist. Ass. 86, 693-698.

Thomas W., Cook R.D. 1989. Assessing influence on regression coefficients in generalized linear models. Biometrika. 76, 741-749.

Thomson. W. R. 1935. On a criterion for the rejection of observations and the distribution and the sample mean in samples of n from a normal universe. Biometrika; 32, 301-310.

Wahba, G. 1990. Spline Model for Observational Data, Siam, Philadelphia Pa.