Journal of Data Science **15**(2017), 275-292

CASE DELETION DIAGNOSTICS IN LIU SEMIPARAMETRIC REGRESSION MODELS

Hadi-Emami[1]

[1]Department of Statistics, University of Zanjan, Zanjan, Iran

Abstract: In semiparametric regression it is of interest to detect anomalous observations that exert an unduly large influence on the parameter's esti-mate and fitted values. Usually the existence of influential observations is complicated by the presence of collinearity. However no method of influ-ence diagnostics available for the possible effects that collinearity can have on the influence of an observation on the estimates of parametric and non-parametric component of semiparametric regression models. In this paper we show when Liu estimators are used to mitigate the effects of collinearity the influence of some observations can be drastically modified. We propose a case deletion formula to detect influential points in Liu estimators of semi-parametric regression models . As an illustrative example a real data set are analysed.

Key words: Bandwidth, Cross validation, Diagnostics, Leverages, Liu esti-mator.

1. Introduction

Diagnostic techniques for the regression model have received great attention in statistical literature. See Cook (1977), Belsley et al. (1989), Walker and Brich (1988) among others. In nonparametric and semiparametric regression models, diagnostic results are quite rare; among them one can refer to Eubank (1985), Silverman (1985), Thomas (1991), and Kim (1996) who studied the basic diag-nostic building blocks such as the residuals, the leverage and the local influence for the choice of smoothing parameter.

In the analysis of influential observations to measure the impact of the *i*th observation, the most common approach is to compute single-case diagnostics with the *i*th case deleted. Since the pioneering work of Cook (1977), case dele-tion diagnostics such as Cook's distance or the likelihood distance has been suc-cessfully applied to various statistical models. Fung et al. (2002), Kim et al.(2001, 2002) and Kim et al. (2002) applied the same case deletion method via

some type of Cook's distance in local polynomial regression and semiparametric regression respectively. In regression analysis, researchers often encounter the problem of multicollinearity. The remedies for the problem of multicollinear-ity depend on the objective of the regression analysis. The multicollinearity is a problem when the primary interest is in the estimation of the parameters in a regression model. In ordinary regression many important biased estimation models have been proposed to deal with multicollinearity; among these the or-dinary ridge regression (RR) models and the Liu estimator (LE) of Liu (1993) are popular. Application of the conventional influence measures in the RR and LE of ordinary linear regression has seen a great surge of research activities during the last two decades or so. It is clear from several articles that appeared in the lit-erature; see, e.g., Walker and Brich (1988), Asar and Erisoglu (2016) and Emami and Emami (2015). However, the LE is generalized to fit the semiparametric regression model for multicollinearity data (see Akdeniz and Akdeniz Duran (2010), Duran et al. (2012) and Duran et al. (2011)). There does not seem to have any work on diagnostics for semiparametric regression models in present of collinearity. Recently, Emami (2015) and Emami (2016) developed influence diagnostics based case deletion and local influence approach for ridge estima-tors and Liu estimators in semiparametric regression models, respectively. In this paper, therefore, we propose a case deletion formula to detect influential points in LE for semiparametric models. We assess the global influence of obser-vations on the LE using the method of case deletion suggested by Walker and Brich (1988). The key to making deletion diagnostics usable is the development of efficient computational formulas, allowing one to obtain the case deletion di-agnostics by making use of basic building blocks, computed only once for the full model. The goal of this paper is to supplement the work of Walker and Brich (1988) with such information and extend some ordinary linear regression influence diagnostics approach to LE in the linear semiparametric models. In the following sections, it is shown graphically how the influence of some case can be modified when LEs are used to reduce the level of collinearity. In section 2 the semiparametric models with LEs are introduced, the relevant notation and some inferential results are also given. Section 3 derives some type of Cook dis-tance and case-deletion formulas for LEs in semiparametric regression models. Statistical properties and motivation of these measures are discussed. In section 4 the proposed methods are illustrated through a simulation study and a real data set. A discussion is given in the section 5.

2. background and definition

Considered the semiparametric regression model

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

where β is a p-vector of regression coefficients, x_i is a p-vector of explanatory variables, t_i is a scalar $(a \leq t_i, \ldots t_n \leq b)$, and $t_i's$ are not all identical, f is a smooth curve and the errors ϵ_i are uncorrelated with zero mean and constant variance σ^2. the model (1) has been used in discussion of many methods, e.g., penalized least square (see Fung et al. (2002)), smoothing spline (see Speckman (1988), Green and Silverman (1994)). In matrix-vector notation, model (1) is written as

$$y = X\beta + f(t) + \epsilon, \tag{2}$$

where $y = (y_1, \ldots, y_n)$, $X' = (x_1, \ldots, x_n)$, $f = (f(t_1), \ldots, f(t_n))'$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. There are several ways of estimating β and f. Let $S \equiv S(t_1, \ldots, t_n)$ be $n \times n$ positive definite smoother matrix and let $\tilde{y} = (I - S)y$ and $\tilde{X} = (I - S)X$, then the estimator of β and f suggested by Speckman (1988) are given by

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

and

$$\hat{f} = S(y - X\hat{\beta}). \tag{4}$$

The estimator of σ^2 is $s^2 = e'e/(n - p)$, where e is the vector of residuals $e = y - \hat{y}$.

2.1. Liu estimators (LEs)

The multicollinearity is a problem when the primary interest is in the estimation of the parameters in a regression model. In the case of multicollinearity we know that when the correlation matrix has one or more small eigenvalues, the estimates of the regression coefficients can be large in absolute value. The least squares estimator performs poorly in the presence of multicollinearity. Some biased estimators have been suggested as a means to improve the accuracy of the parameter estimate in the model when multicollinearity exists. In semiparametric regression models to combat multicollinearity LEs of β and f are defined as

$$b_d = (\tilde{X}'\tilde{X} + I_p)^{-1}(\tilde{X}'\tilde{y} + d\hat{\beta}) \qquad 0 \leq d \leq 1, \tag{5}$$

and

$$\hat{f}_d = f(t, d) = S(y - Xb_d), \tag{6}$$

which are obtained by minimizing the penalizing sum of square function, namely,

$$\|\tilde{y} - \tilde{X}\beta\|^2 + \|d\hat{\beta} - \beta\|^2, \tag{7}$$

where I_p is identity matrix with order p and d is tuning parameter (see Duran et al. (2012)). If $d = 1$, and assume $\tilde{X}'\tilde{X}$ is of full rank then LEs became a two-step estimators of semiparametric model. Let $f(t) \equiv 0$. Then (5) becomes LE of a

linear model. Several methods have been motivated by Liu (1993) to find an estimate for the parameter d in the LE of linear regression. All these methods are similar to those for the choice of k in the ridge regression.

3. Influence Diagnostic in LEs

3.1. Leverage and Residuals

From LE estimator in relations (5) and (6) the vector of fitted values can be written as

$$\begin{aligned}\hat{y}_d &= Xb_d + \hat{f}_d \\ &= (\tilde{H}_d + H_d^*)y \\ &= H_d y,\end{aligned}$$

where $\tilde{H}_d = X(\tilde{X}'\tilde{X} + I_p)^{-1}[I_p + d(\tilde{X}'\tilde{X})^{-1}]\tilde{X}'(I - S)$ and $H_d^* = S(I - \tilde{H}_d)$. $H_d = \{h_{ij,d}\}$ is the hat matrix which is given by

$$H_d = S + (I - S)X(\tilde{X}'\tilde{X} + I_p)^{-1}[I_p + d(\tilde{X}'\tilde{X})^{-1}]\tilde{X}'(I - S).$$

The hat matrix H_d plays the same role as the hat matrix in ordinary least square (OLS). The ith fitted value can be written in term of H_d as $y_i = \sum_{j=1}^{n} h_{ij,d}y_j$, consequently $\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii,d}$. The hat diagonal $h_{ii,d}$ of semiparametric model with LEs can be interpreted as leverage in the same sense as the hat diagonals in semiparametric model with regular estimator. Using the single value decomposition (SVD), the matrix $\tilde{X}$ can be decomposed as $\tilde{X} = UDV$, where the columns of U are orthonormal eigenvectors of $\tilde{X}'\tilde{X}$; the columns of V are orthonormal eigenvectors of $\tilde{X}\tilde{X}'$ and D is a $p \times p$ diagonal matrix containing the squared roots of eigenvalues (γ_j) of matrix $\tilde{X}'\tilde{X}$. The LEs leverage for the ith observation can be written as

$$h_{ii} = s_{ii} + \sum_{r=1}^{n}\sum_{j=1}^{p} \frac{(\gamma_j + d)}{1 + \gamma_j} u_{ij}u_{jr}a_{ri},$$

in which u_{ij} and a_{ij} are the ijth element of matrix U and $I - S$ respectively. Now, the LEs residual vector can be evaluated as

$$\begin{aligned}\hat{e}_d &= y - \hat{y}_d \\ &= y - X\hat{b}_d \\ &= (I_n - H_d)y\end{aligned}$$

3.2. Case Deletion

To consider the effect of dropping the ith case from the data set in a cost-effective

manner requires the use of simple, inexpensive updating formulas. In this sec-tion, we provide such formulas. When properly expressed, the updating for-mulas are remarkably similar to the approximate updating formulas used in parametric linear regression (Asar and Erisoglu (2016)). It is common to com-

pare the estimates $(b_d, \hat{f}_d)$ with the estimates $(b_{d(i)}, \hat{f}_{d(i)})$ which correspond to the case deletion model with the ith case deleted.

Theorem 3.1: The approximate updating formula for $b_{d(i)}$ and $\hat{f}_{d(i)}$ under case deletion are

$$b_{d(i)} = b_d - \left(\frac{\tilde{X}'\tilde{X} + I_p)^{-1}(\tilde{X}'\tilde{X}' + dI_p)(\tilde{X}'\tilde{X})^{-1}\tilde{X}'(I - S)\xi_i e_{d,i}}{1 - h_{ii,d}} \right), \tag{8}$$

$$\hat{f}_{d(i)} = \hat{f}_d - \frac{H_d^* \xi_i e_{d,i}}{1 - h_{ii,d}}, \tag{9}$$

where ξ_i is an $n \times 1$ vector with 1 at ith position and zero elsewhere, e_{di} is the ith element of the LEs residual vector e_d.

Proof: Let $y^* = (y_1^*, \ldots, y_n^*)'$ and

$$y_j{}^* = \begin{cases} y_i^* & j = i \\ y_j & j \neq i \end{cases}$$

where $y_i^* = x_i' b_{d(i)} + \hat{f}_{d(i)}$. Let $\tilde{y}^* \triangle (I - S)y^*$, for any β and smooth curve f, from the definition of $b_{d(i)}$ and $\hat{f}_{d(i)}$ we have

$$\|\tilde{y}^* - \beta\tilde{X}\|^2 + \|d\beta_{(i)} \hat{} - \beta\|^2 = (\tilde{y}_i^* \quad - \tilde{x}_i{}' \beta)^2 + \sum_{j \neq i} (\tilde{y}_j - \tilde{x}_j{}' \beta) + \|d\hat{\beta_{(i)}} - \beta\|^2$$

$$\geq \sum_{j \neq i} (\tilde{y}_j - \tilde{x}_j{}' \beta)^2 + \|d\hat{\beta_{(i)}} - \beta\|^2$$

$$\geq \|\tilde{y}^* - b_{d(i)}\tilde{X}\|^2 + \|d\beta_{(i)} - \hat{} b_{d(i)}\|^2$$

where $\tilde{x}_i{}'$ is ith row of $\tilde{X}$. It follows that $b_{d(i)}$ minimizes $\|$

$$\tilde{y}^* - \tilde{X}\beta\|^2 + \|d\hat{\beta}_{(i)} - \beta\|^2,$$

$\hat{\beta}_{(i)}$ minimizes $\|\tilde{y}_{(i)} - \quad \tilde{X}_{(i)}\beta\|^2$ or approximately minimizes $\|\tilde{y^*} - \beta\|^2$ which $\tilde{X}$ give us $\hat{\beta}_{(i)} = (\tilde{X}'\tilde{X})^{-1} \quad \tilde{X}'\tilde{y}^*$ so that

$$b_{d(i)} = (\tilde{X}\tilde{X}' + \quad I_p)^{-1}(\tilde{X}'\tilde{X}' \quad + dI_p)(\tilde{X}\tilde{X} \quad)^{-1} \tilde{X}'\tilde{y}^*$$

$$= b_d - \left(\tilde{X}'\tilde{X} + I_p)^{-1}(\tilde{X}'\tilde{X}' \quad + dI_p)(\tilde{X}\tilde{X} \quad)^{-1} \tilde{X}'(I - S)\xi_i(y_i - y_i^*) \right)$$

and

$$\hat{f}_{d(i)} = S(y^* - Xb_{d(i)})$$
$$= \hat{f}_d - S[I - X(\tilde{X}'\tilde{X} + I_p)^{-1}(\tilde{X}'\tilde{X}' + dI_p)(\tilde{X}'\tilde{X})^{-1}\tilde{X}'(I - S)]\xi_i(y_i - y_i^*).$$

Also we can verify that

$$\begin{aligned} y_i - y_i^* &= \xi_i(y - Xb_{d(i)} - \hat{f}_{d(i)}) \\ &= \xi_i'(y - H_d y^*) = \xi_i'[y - H_d y + H_d(y - y^*)] \\ &= \xi_i'(y - H_d y) + \xi_i' H_d \xi_i(y_i - y_i^*), \end{aligned}$$

solving for $y_i - y_i^*$ we have

$$y_i - y_i^* = e_{i,d}/(1 - h_{ii,d}),$$

from which Theorem 3.1 follows. If the non-parametric component f is not present in the model, equation (1) reduces to the to the updating formula of Cook distance for ordinary ridge regression (see Walker and Brich (1988)).

3.3. Measuring Influence in LEs

Some version of Cook's distance for the component of semiparametric regression models is suggested by Eubank (1985) and Kim et al. (2001, 2002). An analogous definition of Cook's distance can be given for LEs in semiparametric regression model (1) as follows:

3.3.1. Influence on b_d

At least two versions of Cook's distances can be constructed for LE of β, namely,

$$D_{bi} = \frac{(b_d - b_{d(i)})'X'X(b_d - b_{d(i)})}{p_1 s^2} \tag{10}$$

and

$$D_{bi}^* = \frac{(b_d - b_{d(i)})'K[\tilde{X}'(I - S)(I - S)'\tilde{X}]^{-1}K(b_d - b_{d(i)})}{p_1 s^2}, \tag{11}$$

in which $p_1 = \sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{h}_{ij}^2$, where $\tilde{h}_{ij}$ is ijth element of matrix $\tilde{H}_d$ and $K = (\tilde{X}'\tilde{X} + I)^{-1}(\tilde{X}'\tilde{X} + dI)$, D_{bi}^* is direct generalization of Cook's distance, and D_{bi}^* is based on the fact that $var(\hat{\beta}_k) = K'^{-1}(\tilde{X}'\tilde{X})K'^{-1}$.

The mean squared error is a function of the fitted values and the responses, neither of which depends on individual eigenvalues of $\tilde{X}'\tilde{X}$. Therefore, it is not affected by collinearity. For this reason, the regular estimators of σ^2 [s and $s_{(i)}$] will be used as measures of scale. It would be desirable to be able to write these

measures as functions of leverage and residual. By the theorem we can express D_{bi} and D^*_{bi} as a function of ith residual and leverage as follows:

$$D_{bi} = \frac{[\sum_{j=1}^{n} \tilde{h}^2_{ij}]e^2_{di}}{p_1 s^2 (1-h_{d,ii})^2},$$

$$D^*_{bi} = \frac{h^0_{ii} e^2_{di}}{p_1 s^2 (1-h_{d,ii})^2},$$

in which h^0_{ij} is ith diagonal elements of matrix $(I-S)\tilde{X}'(\tilde{X}'\tilde{X})^{-1}[\tilde{X}'(I-S)(I-S)'\tilde{X}]^{-1}(\tilde{X}'\tilde{X})^{-1}\tilde{X}'(I-S)$.

3.3.2. Influence on $\hat{f}_d$

An influence measure for ith observation on $\hat{f}_d$ may be defined as a type of Cook's distance by

$$D_{fi} = \frac{(\hat{f}_{d(i)} - \hat{f}_d)'(\hat{f}_{d(i)} - \hat{f}_d)}{p_2 s^2}. \tag{12}$$

D_{fi} can be expressed in the form of residuals and leverage's as

$$D_{fi} = \frac{[\sum_{j=1}^{n} h^{*2}_{ij}]e^2_{d,i}}{p_2 s^2 (1-h_{d,ii})^2},$$

where $p_2 = \sum_{i=1}^{n}\sum_{j=1}^{n} h^{2*}_{ij}$ and h^*_{ij} is ijth element of matrix H^*_d.

3.3.3. Influence Measure for $\hat{y}_d$

Among the most popular single-case influence measures is the difference in fit standardized DFFITS (Belsley et al. (1989)), which can be defined for model (1) as

$$DFFITS_d = \frac{x_i'(b_d - b_{d(i)}) + (\hat{f}_d - \hat{f}_{d(i)})}{se(\hat{y}_d)}, \tag{13}$$

the denominator is an estimator of the standard error of the LEs fitted value. It can be shown that $se(\hat{y}_d) = s[\sum_{j=1}^{n} h^2_{d,ij}]$is an estimator of the standard error of the fitted values. Based on the relations in (7) and (8), the relation in (13) can be written as

$$DFFITS_d = \left[\frac{h_{d,ii}}{1-h_{d,ii}}\right] \frac{e_{d,i}}{s[\sum_{j=1}^{n} h^2_{d,ij}]}.$$

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

$$D_{yi} = \frac{(\hat{y}_d - \hat{y}_{d(i)})'(\hat{y}_d - \hat{y}_{d(i)})}{p_3 s^2 (1 - h_{d,ii})^2}, \tag{14}$$

we may express it as

$$D_{yi} = \frac{\sum_{j=1}^{n} h_{d,ij}^2 e_{d,i}^2}{p_3 s^2 (1 - h_{d,ii})^2},$$

where $p_3 = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{d,ij}^2$.

3.4. Relationship between the Cook's distances

Here we shall investigate the relationships between D_{yi}, D_{bi} and D_{fi}. We have

$$\begin{aligned} p_3 s^2 D_{yi} &= (\hat{y}_d - \hat{y}_{d(i)})'(\hat{y}_d - \hat{y}_{d(i)}) \\ &= \{X(b_d - b_{d(i)}) + (\hat{f}_d - \hat{f}_{d(i)})\}'\{X(b_b - b_{d(i)}) + (\hat{f}_d - \hat{f}_{d(i)})\} \\ &= (b_d - b_{d(i)})X'X(b_d - b_{d(i)}) + (\hat{f}_d - \hat{f}_{d(i)})'(\hat{f}_d - \hat{f}_{d(i)}) \\ &\qquad + 2(b_d - b_{d(i)})'X'(\hat{f}_d - \hat{f}_{d(i)}) \\ &= p_1 s^2 D_{bi} + p_2 s^2 D_{fi} + 2(b_d - b_{d(i)})'X'(\hat{f}_d - \hat{f}_{d(i)}) \end{aligned}$$

In the presence of cross product term $2(b_d - b_{d(i)})'X'(\hat{f}_d - \hat{f}_{d(i)})$, D_{yi} might not be large even though both D_{bi} and D_{fi} are large. Conversely D_{yi} could be large even though both D_{bi} and D_{bi} are small. This is reason we should compute the influence of an observation on each estimator separately(see Kim et al. (2001, 2002)).

3.4.1. Reference values

In non-parametric or partial linear regression models, a general guideline for reference values for the Cooks distances is out of reach. After finding proper tuning parameter d, we suggest take the data-specific method based on a bootstrap idea proposed by Kim and Storer (1996). We take this bootstrap idea as follows:

1. Fit the observed data to model (1),
2. Generate n random number $\epsilon_i^* \sim N(0, s^2)$,
3. With known d, generate n pseudo-response $\hat{y}_i^* = \mathbf{x}_i'\hat{b}_d + \hat{f}_d(t_i) + \epsilon_i^*$
4. Based on $(\mathbf{x}_i, t_i, \hat{y}_i^*)$, compute n numbers of the Cook's distances defined in relations (10)-(13).

5. Repeat steps of 1-4 r times, and
6. Obtain the average of r maximum values.

3.5. Selecting the Smoothing Parameter

Both local-polynomial regression and smoothing splines have an adjustable smooth-ing parameter, say λ. This parameter may be selected by visual trial and error, picking a value that balances smoothness against fidelity to the data. More for-mal methods of selecting smoothing parameters typically try to minimize the mean-squared error of the fit or by some form of cross-validation. Here, the cross-validation or the generalized cross-validation, which might be used to es-timate smoothing parameter, can be written as a function of the residuals and LEs leverages by

$$CV(\lambda) = \sum_{i=1}^{n} \{e_{i,d}/1 - h_{ii,d}\}^2,$$

$$GCV(\lambda) = \sum_{i=1}^{n} \{e_{i,d}/1 - \frac{1}{n} tr(H_d)\}^2.$$

4. Numerical Example

4.1. Simulation study

In this section, we will discuss the simulation study to investigate the effect of influence observations in multicollinear data sets. Adopting the model (1.1) we simulate the response from the following model:

$$y = X\beta + f(t) + \epsilon$$

where $X \sim N_{10}(0, \Sigma)$ and the coefficient vector β is the normalized eigenvector corresponding to the largest eigenvalue of $\tilde{X}'\tilde{X}$. We selected a smooth function of the form $f(t) = m_n f_0(t)$ where $f_0 = 4.26\,[exp(-3.25t) - 4exp(-6.5t) + 3exp(-9.7t)]$ with max $|f(t)| = 9$ (i.e $m_n = 9$). The vector t and ϵ generated from $U(0, 1)$ and $N(0, 0.05)$, respectively. Three different set of data corresponding to $\Sigma = \Sigma_1$, $\Sigma = \Sigma_2$ and $\Sigma = \Sigma_3$ are considered to show the weakly, strong and severely collinear between the explanatory variables, respectively. For this, the off-diagonal ele-ments of covariance matrix of Σ_1,Σ_2 and Σ_3 are set to be $cov(x_i, x_j) = 0.7\sigma_{x_i}\sigma_{x_j}$, $cov(x_i, x_j) = 0.9\sigma_{x_i}\sigma_{x_j}$ and $cov(x_i, x_j) = 0.99\sigma_{x_i}\sigma_{x_j}$, $i, j = 1, \ldots, 10$, respectively, where $\sigma_{x_i} = 4i/5$. For each data set, $5th$, $10th$ and $15th$ observations are replaced as influential observations. With given Σ_1,Σ_2 and Σ_3 the samples of 15, 50 and 100 units are produced for 5000 times. D_{yi}, D_{bi} and D_{fi} are calculated for each

data set. Table1 shows the mean and standard error (in parenthesis) of the values of D_{yi}, D_{bi} and D_{fi} . It can be seen that at least 2 influence measures clearly identify the 5*th*, 10*th* and 15*th* observations as influential observations. The percentage of the correct detection of the influential observations by influence mea-sures are given in Table 2.

4.2. Real data

The Longley (1967) data is a data frame with 7 economical variables, $x1$ = GNP implicit price deflator, $x2$ =Gross National Product ,$x3$ = number of people in the armed forces , $x4$ =number of unemployed, $x5$ =Population , $x6$ = Year and y =number of people employed. This data has been used to explain the effect of extreme multicollinearity on the ordinary least square estimation. The scaled condition number (see Belsley et al. (1989), p. 100) of this data set is 43,275. This large value suggests the presence of an unusually high level of collinearity. Cook (1977) applied Cook's distance to this data and found that cases 5, 16, 4, 10, and 15 (in this order) were the most influential observations in OLS. Walker and Brich (1988) analysed the same data to find anomalous observations in ridge regression using the method of case deletion influential measures. They found that cases 16, 10, 4, 15 and 5 (in this order) were the most influential observations in Cook's and *DFFITS* measures. Shi and Wang (1999) detected the five same cases as Walker and Brich (1988) did but with another order, i.e., 10, 4, 15, 16, and 1, using local influence method. Recently Asar and Erisoglu (2016) and Jahufer and Chen (2011) used this data set and applied case deletion formula and local influence respectively on LE in ordinary linear regression. They found the five case i.e., 4, 6, 15 and 16 and found that the order of influence is changed respect to different d. In leverage and residual measures Asar and Erisoglu (2016) found the same influential observations but the order is changed. In this paper we used the same data set to assess the influential observations in LEs by using the method of case deletion influential measures such as Cook's distances, DFFITS and Leverages. For simplicity of the numerical study we consider the following semiparametric model

$$y_i = \beta 1 x_{i1} + \beta 2 x_{i2} + \beta 3 x_{i3} + \beta 4 x_{i4} + \beta 5 x_{i5} + f(x_{i6}) + \epsilon_i. \qquad i = 1, 2, \ldots 16$$

The smoothing spline was used and the tuning parameter λ =0.006 was selected by minimizing the cross-validation criterion. The LE biasing parameter is es-timated for this data set with d = 0.962. Table 1 show the three type Cook's distances, D_{bi}, D_{gi} and D_{yi} for d = 0 and d = 0.962 model respectively. The ref-erence values are evaluated based on the algorithm described in Section 4 with r =100. It can be seen that cases 5, 6, 10, and 15 in this order are most influential points on $\hat{\beta}_d$ and $\hat{f}_d$. In both influential measures the detected influential cases exactly same but only, the influence of observation 15 on $\hat{f}_d$ is not significant.

However, the influence observation on $\hat{y}_d$ are changed in such a way cases 5, 16, 4 and 10 (in this order) are the most significant influential observations on $\hat{y}_d$. We note one should be careful in comparing the three Cooks distances D_{bi}, D_{fi} and D_{yi}. Since different normalizing constants p_1, p_2 and p_3 are used, direct comparison may be misleading. For example, when D_{f5}=0.135 and D_{y5}=0.561, it is not reasonable to conclude that the 5th observation is 4 times more influential on $\hat{y}_d$ than on $\hat{f}_d$. Each Cooks distance should be used and interpreted within the estimator. For example, when D_{f10}=0.462 and D_{f15}=0.191, we may conclude that the 10th observation is three times more influential than the 15th observation on $\hat{f}_d$.

We have also plotted the most influential cases, discovered using (10), (12) and (13), to study the impact of d on these cases against various values of d in Fig.1. The impact of d on the influence of each case is clear. The influential cases are similar in Fig.1(a) and Fig.1 (b) but the order of influence magnitude varies. Specifically, we observe that although case 5 is the most influential, with ref-erence to both (10) and (12), its influence decreases sharply as d increases. In contrast, the influence of case 6 increases slowly as d increases. There is little difference in Fig. (c) the influential observations are 5, 16, 4 and 10 (in this order) and the order of influence along d is unchanged. The impact of d for influence observation on $\hat{y}_d$ is different from $\hat{\beta}_d$ and $\hat{f}_d$. It is seen that the influence of cases 5 and 4 slowly decreases while influence of case 6 is slowly increases as d grows.The influence of the case 10 remains relatively constant in Fig.(a)-(c). DFFITS, leverages ($h_{d,i}$) and residuals are plotted against d in Fig.(3). The influential observations are the same as those in the Cook's measures, but the order of magnitude is changed. In DFFITS and leverage plot, as the plots shows the DFFITS and leverage of case 6 increases as d grows, whereas case 16 and 5 re-mains as the highest point. The residual, on the other hand is more constant for all cases relative to other plot. The joint effect of leverage and residuals is reflected in the change of influence observed in Fig.(1)(c). Fig (3). contain the plot of $\hat{f}_d$ when the influential observations are deleted and when d = 0.962.

5. Summary

In this article, we discussed case omission measures for semiparametric models with the LE estimators. In general linear model Belsley et al. (1989) noted that biased estimators are used to reduce the affect of multicollinearity and that the influence of some cases can be modified. Based on this fact, they suggested that multicollinearity should be controlled before attempting to measure influence. In this paper, we show that, in semiparametric regression models when LE is used, the influence of some case can be changed along d. We found that when the value of d is determined, influence measures should be computed for that d. The main advantage of deletion formulas derived in theorem 1 is that, the

estimator does not have to be computed every time a case is deleted. For a value of d all of the elements in (10)-(13) are readily available from a single run of LE. Also, these measures, based on deletion formulas, are particularly helpful for large data sets. Furthermore,the deletion formulas provide computationally inexpensive influence measures for LE in semiparametric regression models.

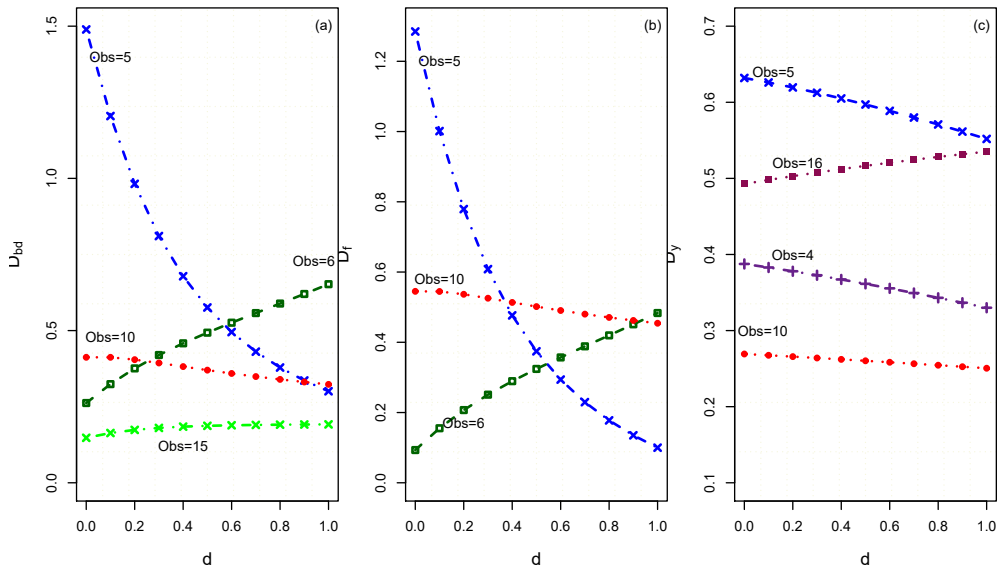

Figure 1: (a): Plot of D_{bi} against d, (b): Plot of D_{fi} against d, (c): Plot of D_{yi} against d

Table 1: The three most influential observations according to the D_{yi}, D_{bi} and D_{fi} in the simulation study

Σ	n	$Obs.$	D_{yi}	D_{bi}	D_{fi}
Σ_1	15	5	0.0943(0.0281)	0.1941(0.0105)	0.0605(0.1061)
		10	0.0328(0.0176)	0.0997(0.0755)	0.0710(0.1100)
		15	0.1003(0.0139)	0.1203(0.0182)	0.09114(0.0152)
		other	$\leq$ .02726(0.0301)	$\leq$ .04174(0.0392)	$\leq$ .0534(0.0608)
	50	5	0.1253(0.0150)	0.1318(0.0194)	0.1967(0.0167)
		10	0.1827(0.0121)	0.101(0.0110)	0.0979(0.0182)
		15	0.228(0.0151)	0.2027(0.0141)	0.169(0.0315)
		other	$\leq$ .0493(0.0107)	$\leq$ .0433(0.0100)	$\leq$ .0735(0.0089)
	100	5	0.5010(0.0081)	0.4010(0.0200)	0.2101(0.0092)
		10	0.4490(0.0087)	0.2783(0.0096)	0.2079(0.0099)
		15	0.3641(0.0078)	0.5502(0.0097)	0.3044(0.0105)
		other	$\leq$ .0140(0.0103)	$\leq$ .0162(0.0091)	$\leq$ .0094(0.0131)
Σ_2	15	5	0.129(0.0028)	0.491(0.0088)	0.1930(0.0087)
		10	0.3914(0.0044)	0.3152(0.0067)	0.2671(0.0061)
		15	0.2149(0.0073)	0.3861(0.0092)	0.2918(0.0092)
		other	$\leq$ .0621(0.0073)	$\leq$ .0346(0.0290)	$\leq$ .068(0.0053)
	50	5	0.3461(0.0055)	0.2183(0.0083)	0.2192(0.0099)
		10	0.3410(0.0098)	0.2093(0.0074)	0.1901(0.0039)
		15	0.2990(0.0048)	0.2088(0.0081)	0.1011(0.0102)
		other	$\leq$ .0194(0.0097)	$\leq$ .0504(0.0082)	$\leq$ .0089(0.0069)
	100	5	0.314(0.0064)	0.2100(0.0067)	0.1547(0.0066)
		10	0.2910(0.0083)	0.3094(0.0076)	0.2819(0.0086)
		15	0.2116(0.0076)	0.1512(0.0091)	0.1989(0.0043)
		other	$\leq$ .0069(0.0081)	$\leq$ .0095(0.0056)	$\leq$ .0080(0.0061)
Σ_3	15	5	0.1319(0.0027)	0.2852(0.0089)	0.1810(0.0071)
		10	0.2671(0.0082)	0.1845(0.0086)	0.1906(0.0089)
		15	0.2354(0.0028)	0.1839(0.0086)	0.251(0.0059)
		other	$\leq$ .0094(0.0079)	$\leq$ .0154(0.0081)	$\leq$ .063(0.0097)
	50	5	0.1328(0.0080)	0.1715(0.0061)	0.3014(0.010)
		10	0.5951(0.0064)	0.2627(0.0074)	0.2199(0.0076)
		15	0.1498(0.0088)	0.2691(0.0026)	0.2990(0.0094)
		other	$\leq$ .0061(0.0084)	$\leq$ .0069(0.0083)	$\leq$ .0076(0.0099)
	100	5	0.2891(0.0081)	0.2816(0.0099)	0.1107(0.0099)
		10	0446(0.0078)	0.2638(0.0066)	0.2091(0.0058
		15	0.2951(0.0075)	0.2804(0.0029)	0.2769(0.0039)
		other	$\leq$ .0221(0.0049)	$\leq$ .0086(0.0029)	$\leq$ .0161(0.0097)

Table 2: The percentage of correct detection in the simulation study with $n = 50$

Σ	% of outliers	D_{yi}	D_{bi}	D_{fi}
Σ_1	5%	100	100	99
	10%	99	98	97
	25%	98	99	97
Σ_2	5%	98	99	98
	10%	97	97	96
	25%	96	96	96
Σ_3	5%	97	97	95
	10%	96	96	94
	25%	95	96	94

Table 3: The 6 most influential observations: Longley data.Values with $*$ indicates the cases with values larger than the reference values.

	d=0			d=0.962		
Case	D_{bi}	D_{fi}	D_{yi}	D_{bi}	D_{fi}	D_{yi}
4	0.005	0.005	0.378*	0.002	0.003	0.337*
5	1.489*	1.284*	0.631*	0.336*	0.135*	0.561*
6	0.262*	0.017	0.093	0.621*	0.451*	0.091
10	0.412*	0.545*	0.269*	0.331*	0.462*	0.250*
15	0.147*	0.190*	0.088	0.191*	0.191*	0.096
16	0.042	0.047	0.492*	0.054	0.059	0.532*

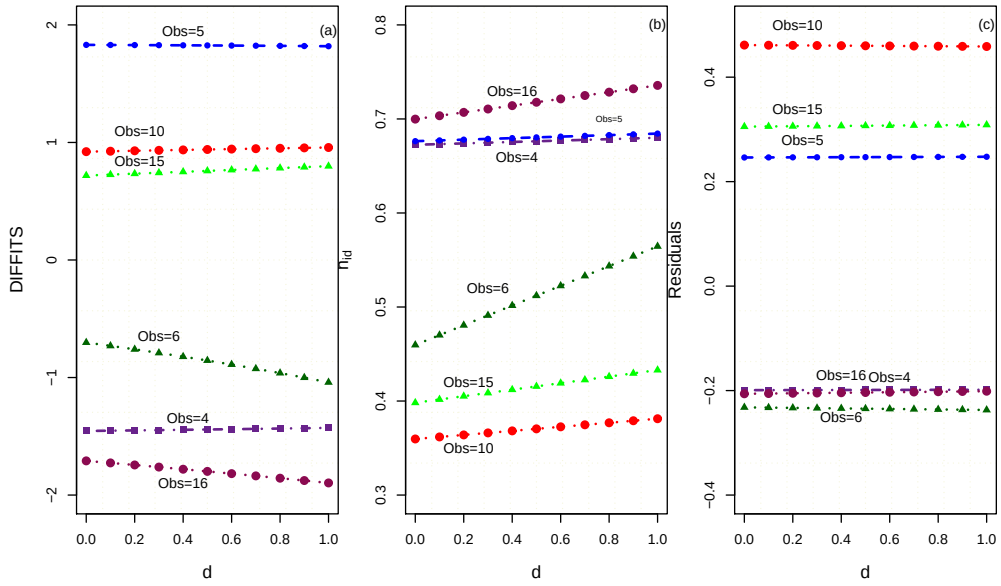

Figure 2: (a): Plot of DFFITS against d, (b): plot of leverages against d, (b): Plot of residuals against d

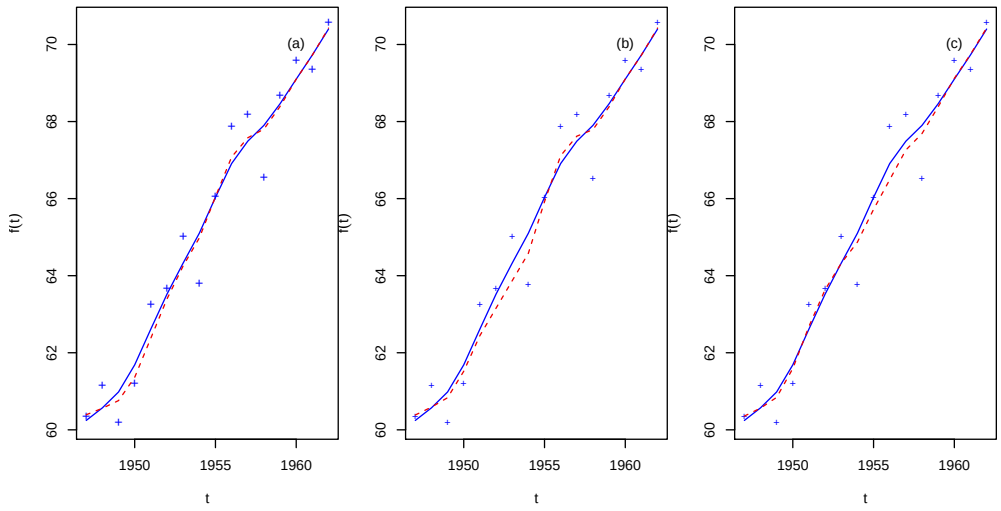

Figure 3: Plot of $\hat{f}_d$ (solid line) when case 5 is deleted (dashed line) (a), when case 10 is deleted (b),when case 6 is deleted (c).

Acknowledgements

We are grateful to the associate editor and two referees for their helpful com-ments and suggestions on earlier drafts of the paper.

References

1. Akdeniz,. F. and Akdeniz, Duran. E.(2010). Liu-type estimator in semiparametric segression models. Stat Comput Sim. 80, 853-871.Asar, Y. and Erisoglu M. (2016). Influence Diagnostics in Two-Parameter Ridge

2. Regression . Journal of Data Science 14,: 33-52.

3. Belsley, D.A., Kuh, E. and Welsch, R.E. (1989). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.

4. New York : Wiley. Cook, R.D. (1977). Detection of influential observations in linear regression. Technometrics. 19,:15–18.

5. Duran, E., Hardel, W.K. and Osipenko, M.(2012). Difference based ridge and Liu type estimators in semiparametric regression models. Multivariate. 105,164-175.

6. Duran, E., Akdiniz , F. and Hu, H.(2011). Efficiency of a Liu-type estimator in semiparametric regression models.Comput. Appl. Math. 235, 1418–1428.

7. Emami,. Hand Emami, M.(2015). New Influence diagnostic in ridge regression. Applied Statistics. 43,1-16.

8. Emami, H.(2015). Influence diagnostics in ridge semiparametric regression models. Stat & Prob Lett. 105, 106–115.

9. Emami, H.(2016). Local influence for Liu estimators in semiparametric linear models. Stat Paper. DOI: 10.1007/s00362-016-0775-6

10. Eubank, R.L.(1985). Diagnostics for smoothing splines. Roy. Statist. Soc. ;B47,:332-341.

11. ified ridge regression. Journal of Data Science. 9, 359-372.

12. Fung, W.K., Zhu, Z.Y., Wei, B.C. and He, X.(2002). Influence diagnostics and outlier tests for semiparametric mixed models. Roy. Statist. Soc.. B47, 332-341.

13. Green, P.J. and Silverman, B.W.(1994). Non-parametric Regression and Generalized Linear Models. London: Chapman and Hall.

14. Jahufer, A. and Chen., J.(2012).Measuring local influential observations in modified ridge regression. Journal of Data Science. 9, 359-372.

15. Kim, C.(1996). Cook’s distance in spline smoothing. Statist Probab Lett. 31, 139-144.

16. Kim, C., Lee, Y. and Park, B.U., (2001). Cook’s distance in local polynomial regression. Statist. Probab. Lett. 54, 33-40.

17. Kim, C., Lee, Y. and Park, B.U.(2002). Influence diagnostics in semiparametric regression models. Statist. Probab. Lett. 60,49-58.

18. Kim, C. and Storer, B.E.(1996). Reference values for Cook’s distance. Comm. Statist. Sim. 25, 691-709.

19. Kim, C. and Kim, W.(1998). Some diagnostics results in non-parametric density estimation.Comm. Statist. Theor. Meth. 27, 291-303.
20. Liu, K.(1993). A new class of biased estimate in linear regression. Commun. Statist. Theor. Meth. 22, 393-402.
21. Longley, J.W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. J. Amer. Statist. Assoc. 62,819-84.
22. Shi, L. and Wang, X.(1999). Local influence in ridge regression.Computat. Statist. Data Anal. 31,341-353.
23. Silverman, B.W.(1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). Roy. Statist. Soc. B47: 1-52.
24. Speckman, P.E.(1985). Regression analysis for partially linear models. Roy. Statist. Soc. B50,413-436.
25. Thomas , W.(1991). Influence diagnostics for the cross-validated smoothing parameter in spline smoothing. J. Amer. Statist.Assoc. 86, 693-698.
26. Walker, E. and Birch, J.B. (1988). Influence measures in ridge regression.

Technometrics.30: 221-227.
Hadi-Emami
Department of Statistics
University of Zanjan
Zanjan, Iran
Email: h.emami@znu.ac.ir