Statistical Papers
https://doi.org/10.1007/s00362-021-01279-4

REGULAR ARTICLE

# Local influence diagnostics with forward search in regression analysis

**Reiko Aoki**[1] · **Juan P. M. Bustamante**[1] · **Gilberto A. Paula**[2]

Received: 19 February 2021 / Revised: 24 September 2021 / Accepted: 25 November 2021


## Abstract

Regression analysis is one of the most widely used statistical techniques. It is well known that the least squares estimates is sensitive to atypical and/or influential observations. Many methodologies were proposed to detect influential observations considering case deletion (global influence). On the other hand, Cook (J R Stat Soc Ser B 48(2):133–169, 1986) developed a general and powerful methodology to obtain a group of observations that might be jointly influential considering the local influence. However, these techniques may fail to detect masked influential observations. In this paper, we propose a methodology to detect masked influential observations in a local influence framework considering the forward search (Atkinson and Riani, Robust diagnostic regression analysis, Springer, New York, 2000). The usefulness of the proposed methodology is illustrated with data sets which were previously analyzed in the literature to detect outliers and/or influential observations. Masked influential observations were successfully identified in these studies. The proposed methodology may be used in any model where the local influence analysis (Cook 1986) is appropriate.

**Keywords** Regression model · Masked observations · outliers · Influential observations · Likelihood displacement

**Mathematics Subject Classification** 62J20

✉ Reiko Aoki
reiko@icmc.usp.br

Juan P. M. Bustamante
juanesta@icmc.usp.br

Gilberto A. Paula
giapaula@ime.usp.br

[1] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP, Brazil

[2] Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, SP, Brazil

Published online: 28 January 2022

## 1 Introduction

Influence diagnostic is an important step in the analysis of a statistical model. It is well known that parameter estimation and variable selection are very sensitive to the presence of outlying or influential observations in regression analysis (Atkinson and Riani 2000; Montgomery et al. 2011). There is an extensive literature on detecting outliers and influential observations in linear regression, such as DFBETA, DFFIT (Belsley et al. 1980) and the Cook's distance (Cook 1977). These techniques were quickly assimilated being extended to several classes of models and many works deal with the global influence scheme. The global influence consists in the evaluation of the impact of removing one or more observations in the data analysis. Mavridis and Moustaki (2009) described two kinds of errors that may occur in the process of detecting outliers. The first one is the masking effect and it occurs when an outlier is undetected because of the presence of a cluster of outliers. The second one is the swamping effect which occurs when an observation is incorrectly identified as an outlier. The traditional deletion methods may fail to detect multiple outliers as they can be affected by own observations that should be identified (masking effect). Cook and Weisberg (1982, p. 31) pointed out that residuals with opposite sign can mask each other in a way that none appears as atypical data in the diagnostic methods considering residuals. Chatterjee and Hadi (1988) discussed methods to detect observations which are jointly influential, but individually may not be considered atypical, points of leverage or influential observations, which is considered as masking effect. Hadi (1992) dealt with the identification of multiple atypical data. Hadi and Simonoff (1993) proposed methods that are less susceptible to masking problems.

Atkinson and Riani (2000) developed a method based on the articles of Hadi (1992) and Hadi and Simonoff (1993), which was called forward search algorithm to detect multiple masked outliers in regression models and determine their effect on inferences of the fitted model. The methodology proposed by Atkinson and Riani (2000) consists in obtaining a small subset of the data free of outliers and gradually increment the subset until all the observations are included. The parameters of the model are estimated from the subset, but the criterion to enter into the subset is applied to all the observations. They considered the least-squares estimates and the coefficients estimated in this way were used to evaluate the residuals of all the observations. Then the least median of squares of the observational residuals were used to select the initial subset with $p + 1$ (number of parameters) observations. Thereafter, the $p + 2$ observations with the smallest squared residuals based on a fit of $p + 1$ observations are chosen to compose the next subset.

As the subset size increases, the evolution of residuals, parameter estimates and inferences are monitored and the results are presented as forward plots which show the evolution of the quantities of interest as a function of the size of the subset. In the first steps, the forward search algorithm usually avoids the inclusion of atypical observations, but the choice of the starting subsets does not dramatically influence the search and if an outlier is included in the earlier steps, soon it is disregarded not affecting the final steps. In the last iteration, we have the least-squares estimate of the parameters based on the whole data set.

Many works that extend the preliminary forward search based algorithm were developed in the last decades. Atkinson et al. (2018) uses the forward search to cluster multivariate data, Cerioli et al. (2019) combines the cluster analysis with robust estimation considering the forward search, Riani et al. (2014) and Cerioli et al. (2018) monitor robust high-breakdown procedures, Riani et al. (2019) comments on Galeano and Peña (2019) considering the forward search and Grané et al. (2021) combines forward search distance-based algorithm with robust clustering to visualize mixed data, for example.

On the other hand, instead of deleting cases, Cook (1986) proposed a powerful and general methodology, the local influence analysis, to assess the effect of a minor perturbation in the model or in the data considering the normal curvature of an influence graph based on the likelihood displacement. This methodology can identify a group of observations that are locally influential. It was quickly disseminated and there are numerous applications of this diagnostic method in diverse areas, as can be seen in Beckman et al. (1987), Lawrance (1988), Thomas and Cook (1990), Escobar and Meeker Jr (1992), Paula (1993), Labra et al. (2007), Russo et al. (2009), Russo et al. (2012) and Zhu et al. (2016), for example.

However, the procedure may fail to identify masked influential observations and as far as we know the detection of masked influential observations in local influence analysis Cook (1986) with the use of the forward search has not been addressed yet.

The contribution of this paper is to propose a methodology to detect masked influential observations in a local influence framework considering the forward search algorithm (LIFS). To ease interpretation, the proposed methodology was applied to linear regression models, however it can be applied directly to more complex models as measurement error models or random effects models, for example, and also in univariate and multivariate models, as the methodology may be applied to any model with a smooth likelihood function Cook (1986). In addition, with the forward plot it is possible to see how the influence of each observation changes as the number of elements in the subset increases.

We analyzed data sets which were previously analyzed in the literature to show the usefulness of the proposed methodology. The first and second data sets are rat data and geese data, respectively, which were reported in Weisberg (1980) and analyzed in Cook (1986). The third data set is the bank data, which was considered in Riani et al. (2014) and the fourth data set, customer data, was analyzed in Neter et al. (1996). The first data set contain masked observation considering the local influence analysis, which can easily be detected using the LIFS algorithm. In the second and third data set, there are some influential observations which can easily be detected. The fifth data set is a simulated one to see the effect of a dataset with high contamination rate.

To motivate the proposed methodology, the LIFS forward plot for the rat data set and the geese data set considering the case weight perturbation scheme are presented in Figs. 1 and 2 , respectively. The details of the methodology and more complete illustration with these data sets, analyzed in Cook (1986), can be found in Sects. 2 and 3, respectively. Also, Cook (1986) considered that the error variance $\sigma^2$ is known and the value of $\sigma^2$ was replaced by its estimated value. Here, we consider $\sigma^2$ as unknown parameter in our analysis.

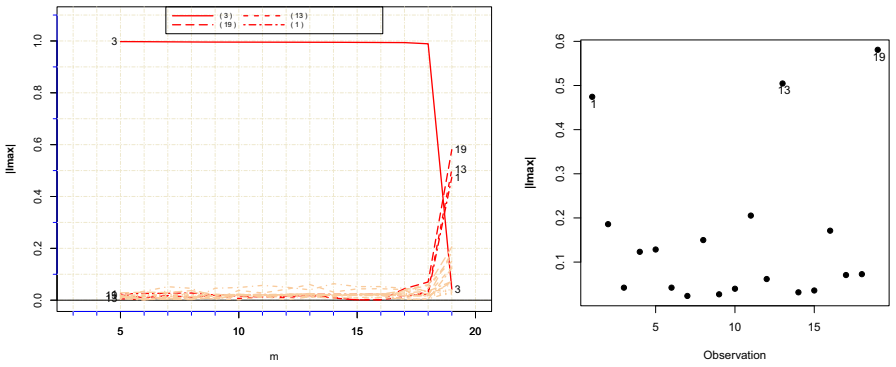


**Fig. 1** Rat data: case-weight perturbation scheme. LIFS forward plot (left) and index plot of |**lmax**| (right)

The left and right hand panel of Fig. 1 shows, respectively, the forward plot of |**lmax**| considering the LIFS algorithm and the index plot of |**lmax**| (Cook 1986) for the rat data. This data set consists of three explanatory variables with $X_1$ representing the body weight, $X_2$ representing the liver weight and $X_3$ the relative dose. A certain amount of the drug was given to each rat and after an amount of time, each rat was sacrificed and the percent of the dose in the liver ($Y$) was determined. The actual dose that each rat received was related to their weight, as liver weight is strongly related to body weight and also large livers would absorb more dose (Weisberg 1980).

The left hand panel of Fig. 1 shows that observation 3 stands out for almost the entire evolution of the LIFS algorithm as an influential observation. In the last iteration observation 3 enters into the search, which makes a steep drop of the observation 3 from the penultimate iteration to the last iteration. Consequently, observation 3 is masked in the last iteration.

Considering Cook (1986) and the right hand panel of Fig. 1 observation 19, followed by observations 13 and 1 stand out. It is not possible to detect that observation 3 might be influential. Observation 19 was the one which had the second greatest absorption by the liver. However, it was the rat with the second lower liver weight. Furthermore, observation 13 was the one that had the least drug absorbtion and observation 1 was the one that among those that received 0.88 mg of the dose (four rats), it had the greatest absorbtion of the dose and the lowest liver weight. According to this index plot of |**lmax**|, we conclude that the group of observations (1, 13, 19) might be jointly influential.

Figure 2 (top) shows the forward plot of |**lmax**| considering the LIFS algorithm and the geese data, where the response variable $Y$ represents the number of birds in the flock counted based on a aerial photograph and $X$ represents the estimated number of geese in the flock made by the observer 1. In this case, the methodology shows that observation 29 is the most influential one, followed by the observations 28 and 41 and there is no masked observations (see Sect. 2 for details). The left hand panel (bottom) of Fig. 2 shows the index plot of |**lmax**| (Cook 1986). Clearly, observations 29, 28 and 41 stand out as influential observations. Also, the scatter plot of the data can be found in the right hand panel (bottom).

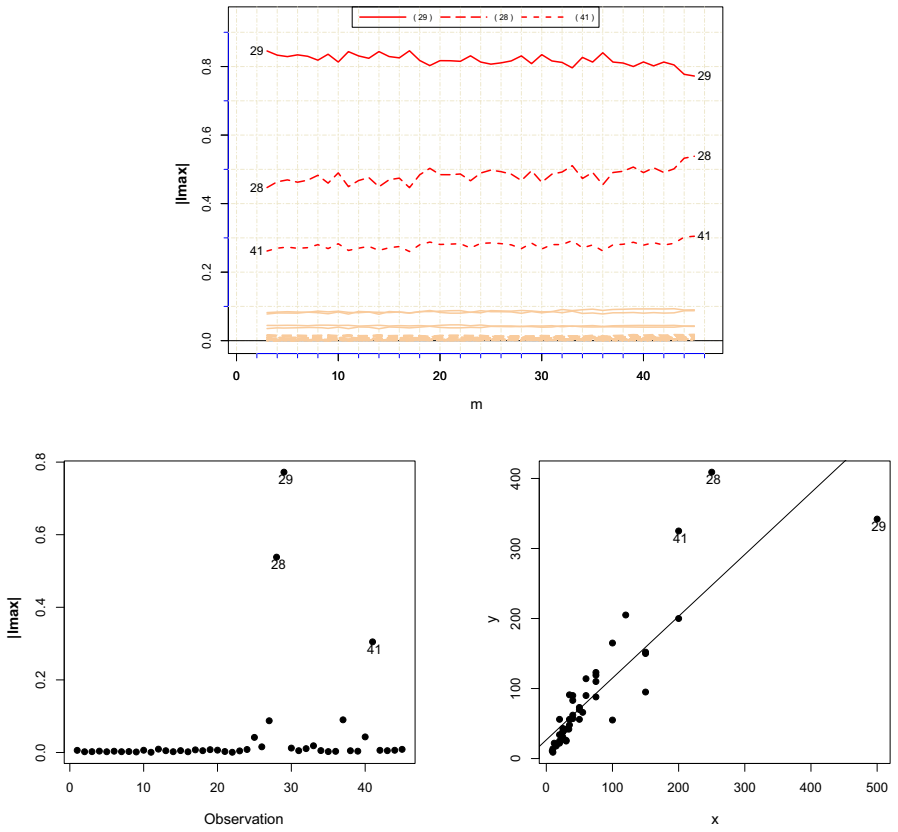


**Fig. 2** Geese data: case-weight perturbation scheme. LIFS forward plot (top) and index plot of |**lmax**| (bottom left). Scatter plot (bottom right)

In Sect. 3 these data sets and also the bank data, the customer data and the simulated data set will be analyzed in more detail with the use of the LIFS algorithm.

The paper is organized as follows. In Sect. 2 we review the local influence technique (Cook 1986) and introduce the LIFS algorithm. Section 3 presents the LIFS algorithm applied to the four data sets previously analyzed in the literature and a simulated data set to show the usefulness of the proposed methodology. In the last section, we discuss the obtained results.

## 2 Local influence and LIFS algorithm

Let

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y}_{(n\times 1)}$ denotes the vector of the dependent variable, $\boldsymbol{X}_{(n\times p)}$ is the model matrix of explanatory variable values, $\boldsymbol{\beta}_{(p\times 1)} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$ denotes the vector of

unknown parameters and $\boldsymbol{\epsilon}_{(n\times 1)}$ the vector of random errors, with $\epsilon_i \overset{ind}{\sim} N(0, \sigma^2)$, $i = 1, \ldots, n$, and $\sigma^2$ unknown.

We denote by $\ell(\boldsymbol{\theta})$ and $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$, respectively, the log-likelihood function of the postulated model and perturbed model, where $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$ is twice continuously differentiable in $(\boldsymbol{\theta}^T, \boldsymbol{\omega}^T)^T$, with $\boldsymbol{\omega} = (w_1, w_2, \cdots, w_q)^T$ representing the perturbation vector in an open subset $\Omega$ of $\mathbb{R}^q$, whereas $\boldsymbol{\theta}_{(p+1)\times 1}$ is the vector of unknown parameters. $\boldsymbol{\omega}_0$ denotes the vector of no perturbation, such that $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}_o) = \ell(\boldsymbol{\theta})$. $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ are, respectively, the maximum likelihood estimator from the postulated and perturbed models.

Cook (1986) proposed an influence measure, the likelihood displacement

$$LD(\boldsymbol{\omega}) = 2\{\ell(\widehat{\boldsymbol{\theta}}) - \ell(\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})\}$$

to assess the influence of varying $\boldsymbol{\omega}$ in $\Omega$. The graph of $\alpha(\boldsymbol{\omega}) = \begin{pmatrix} \boldsymbol{\omega} \\ LD(\boldsymbol{\omega}) \end{pmatrix}$ is named the influence graph whereas the normal curvature $C_{\boldsymbol{l}}$ of the lifted line in the direction $\boldsymbol{l}$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ can be written as $C_{\boldsymbol{l}} = 2|\boldsymbol{l}^T \boldsymbol{\Delta}^T \ddot{\boldsymbol{L}}^{-1} \boldsymbol{\Delta} \boldsymbol{l}|$, where $\ddot{\boldsymbol{L}} = \left.\dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$ with $\boldsymbol{\Delta} = \left.\dfrac{\partial^2 \ell(\boldsymbol{\theta} \mid \boldsymbol{\omega})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\omega}^T}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}, \boldsymbol{\omega}=\boldsymbol{\omega}_0}$ and $\|\boldsymbol{l}\| = 1$.

Cook (1986) suggested the use of **lmax** which is the eigenvector associated with $C_{max}$ that corresponds to the maximum absolute eigenvalue of $\ddot{\boldsymbol{F}} = \boldsymbol{\Delta}^T \ddot{\boldsymbol{L}}^{-1} \boldsymbol{\Delta}$.

For more details related to the local influence analysis, see Cook (1986).

Next, we propose a methodology based on the forward search algorithm and the local influence analysis to deal with the problem of masking effects.

### 2.1 LIFS algorithm

Many methods developed to detect outliers consist in the division of the data in a clean subset and a subset with outliers/influential observations. Atkinson and Riani (2000) developed a general methodology to obtain multiple masked outliers. The main concept is to extract from the dataset with $n$ observations, a small subset of size $m$ free of outliers and estimate the unknown parameters. If the model contains $p$ parameters, they suggested to start the algorithm with a subset of size $m = p$. If $\binom{n}{p}$ is too large, it was suggested to use some large number of subsets, for example 1000. They used the least-squares estimates and then the least median of squares of the observational residuals were applied to select the subset. In the next iteration it was considered subsets of size $m + 1$ and the procedure is repeated until all the observations are included.

In this work we propose the use of the local influence analysis to assess the effect of minor perturbation in the model/data set considering the forward search algorithm. Cook (1986) suggested the use of the maximum curvature $C_{max}$ and the associated eigenvector **lmax** to study the behavior of the influence graph, due to the fact that this is the direction which gives the greatest local change in the likelihood displacement. One way to find out the influential observations is to plot the elements of the

**lmax** and if the $ith$ element or a group of elements of the **lmax** is relatively large, this is an indication that the observation or the group of observations are relatively influential.

The local influence with forward search (LIFS) algorithm starts by choosing a subset of size $m = p + 1$. In order to find the initial subset, we start sampling subsets of size $m = p + 1$ from the data set according to Atkinson and Riani (2000) and estimate the unknown parameters considering the maximum likelihood estimator. Notice that in the case of model (1) the least-squares estimator and the maximum likelihood estimator are the same. If we have sampled $K$ subsets, we will have a set of $K$ maximum likelihood estimates (MLE) corresponding to each subset.

The next step consists in computing the vector **lmax** considering the whole data set, but using the MLE obtained for each of the $K$ subsets of size $m = p + 1$ obtained in the first step. It means that, for each subset, the values of the MLE are obtained considering only the elements of the subset and then considering the whole data set the elements of the vector **lmax** are obtained. So, we will obtain $K$ vectors **lmax** and the subset corresponding to the least median of the **lmax**'s will be chosen in the first step.

The forward search moves to the dimension $m + 1$ and all the steps described so far are repeated but now considering subsets of size $m + 1$. Gradually, the number of observations used in the fit are incremented until all the $n$ observations are used to fit the model.

In the last step of the LIFS algorithm we have only one set left, which is the whole data set. It means that we will derive the usual **lmax** attained considering the whole data set.

The results are presented as forward plots which show the evolution of the vectors **lmax** as a function of the subset size. Other quantities of interest can also be presented as forward plots, such as the estimated value of the parameters.

In order to better illustrate the LIFS algorithm, we are going to describe the methodology step by step considering a simple regression model. In this case, $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)^T$ and the number of parameters is $p + 1 = 3$. We assume that the data set has $n = 5$ observations, $(X_i, Y_i)$, $i = 1, \cdots, 5$ and also that the perturbation scheme was already defined, as well as, the obtention of all the necessary matrices, so that it is possible to obtain the vector **lmax**.

– Step 1: start the LIFS algorithm with all possible subsets of size $m = p + 1 = 3$ from the data set ($C_m^n = 10$).

| Number of the subset | $k$= 1 | $k$=2 | $\cdots$ | $k$=9 | $k$=10 |
|---|---|---|---|---|---|
| Elements in the subset | $(X_1, Y_1)$<br>$(X_2, Y_2)$<br>$(X_3, Y_3)$ | $(X_1, Y_1)$<br>$(X_2, Y_2)$<br>$(X_4, Y_4)$ | $\cdots$<br>$\cdots$<br>$\cdots$ | $(X_2, Y_2)$<br>$(X_3, Y_3)$<br>$(X_5, Y_5)$ | $(X_3, Y_3)$<br>$(X_4, Y_4)$<br>$(X_5, Y_5)$ |
| MLE of the $k$-$th$ subset | $\hat{\boldsymbol{\theta}}_1$ | $\hat{\boldsymbol{\theta}}_2$ | $\cdots$ | $\hat{\boldsymbol{\theta}}_9$ | $\hat{\boldsymbol{\theta}}_{10}$ |
| **lmax** obtained with the whole data set, but with the MLE of the $k$-$th$ subset ($\hat{\boldsymbol{\theta}}_k$) | $\mathbf{lmax}_1$ | $\mathbf{lmax}_2$ | $\cdots$ | $\mathbf{lmax}_9$ | $\mathbf{lmax}_{10}$ |
| | $lmax_{1_1}$<br>$lmax_{1_2}$<br>$lmax_{1_{med}}$<br>$lmax_{1_4}$<br>$lmax_{1_5}$ | $lmax_{2_1}$<br>$lmax_{2_2}$<br>$lmax_{2_{med}}$<br>$lmax_{2_4}$<br>$lmax_{2_5}$ | $\cdots$<br>$\cdots$<br>$\cdots$<br>$\cdots$<br>$\cdots$ | $lmax_{9_1}$<br>$lmax_{9_2}$<br>$lmax_{9_{med}}$<br>$lmax_{9_4}$<br>$lmax_{9_5}$ | $lmax_{10_1}$<br>$lmax_{10_2}$<br>$lmax_{10_{med}}$<br>$lmax_{10_4}$<br>$lmax_{10_5}$ |

- Determine $l_m^* = \min(lmax_{1_{med}}, lmax_{2_{med}}, \cdots, lmax_{9_{med}}, lmax_{10_{med}})$;
- If the vector **lmax** corresponding to $l_m^*$ is $\mathbf{lmax}_4$, for instance, then the vector $\mathbf{lmax}_4$ is the **lmax** obtained in the first iteration for $m = 3$;
- Other quantities of interest, such as $\hat{\boldsymbol{\theta}}_4$ can also be used to detect masked influential observations.

$l_m^*$

– Step 2: move to the next iteration ($m = 4$) and repeat Step 1 with all possible subsets of size $m = 4$ from the data set ($C_m^n = 5$).
– Step 3: move to the next iteration ($m = n = 5$) and repeat Step 1 with all possible subsets of size $m = 5$ from the data set ($C_m^n = 1$).

| Number of the subset | $k$= 1 (whole data set) |
|---|---|
| Elements in the subset | $(X_1, Y_1)$<br>$(X_2, Y_2)$<br>$(X_3, Y_3)$<br>$(X_4, Y_4)$<br>$(X_5, Y_5)$ |
| MLE of the $k$-$th$ subset | $\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\theta}}$ |
| **lmax** obtained with the whole data set, but with the MLE of the $k$-$th$ subset ($\hat{\boldsymbol{\theta}}_k$) | $\mathbf{lmax}_1 = \mathbf{lmax}$ |
| | $lmax_{1_1} = lmax_1$<br>$lmax_{1_2} = lmax_2$<br>$lmax_{1_{med}} = lmax_{med}$<br>$lmax_{1_4} = lmax_4$<br>$lmax_{1_5} = lmax_5$ |

- In the last step, it is simply obtained the usual vector **lmax**.

$l_m^*$

In order to construct the forward plots considering the LIFS algorithm we will consider the model defined in (1) for the first three data sets and the simulated data set with four perturbation schemes.

Case-weight perturbation scheme: in this case, the perturbed log-likelihood function is given by $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^{n} \omega_i \ell(\boldsymbol{\theta})$ and the corresponding matrix $\boldsymbol{\Delta}$ can be written as

$$\boldsymbol{\Delta} = \left( \frac{\boldsymbol{X}^T D(\boldsymbol{e})}{\hat{\sigma}^2}, \frac{\boldsymbol{e}_{sq}^T}{2\hat{\sigma}^4} - \frac{\mathbf{1}_n^T}{2\hat{\sigma}^2} \right)^T,$$

where $\boldsymbol{e} = (e_1, \ldots, e_n)^T$ is the residual vector obtained when $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = \mathbf{1}$ with $\boldsymbol{D}(\boldsymbol{e})$ denoting the diagonal matrix, $\boldsymbol{e}_{sq} = \boldsymbol{e} \odot \boldsymbol{e} = (e_1^2, \ldots, e_n^2)^T$ and $\mathbf{1}_n$ the vector composed by $n$ ones.

Explanatory variable perturbation scheme: the perturbed covariate matrix is given by $\boldsymbol{X}_\omega = \boldsymbol{X} + \boldsymbol{W}\boldsymbol{S}$, where $\boldsymbol{W}_{(n\times p)}$ and $\boldsymbol{S} = \boldsymbol{D}(\boldsymbol{s})$ with $\boldsymbol{s} = (s_0, s_1, \ldots, s_{p-1})^T$ denote, respectively, the perturbation matrix and the scale matrix, with $s_k$ denoting the scale factors accounting for the different metric units related to the columns of $\boldsymbol{X}$, $k = 1, \ldots, p-1$, and $s_0 = 0$. Cook (1986) presented the matrix $\boldsymbol{\Delta}$ when the error variance $\sigma^2$ is known. In our case, the matrix $\boldsymbol{\Delta}$ has dimension $(p+1) \times n(p+1)$ and can be partitioned as $\boldsymbol{\Delta} = (\boldsymbol{\Delta}_0, \boldsymbol{\Delta}_1, \ldots, \boldsymbol{\Delta}_p)$ with the elements $\boldsymbol{\Delta}_k$ being given by

$$\boldsymbol{\Delta}_k = \left( \frac{s_k(\boldsymbol{b}_k \boldsymbol{e}^T - \hat{\beta}_k \boldsymbol{X}^T)}{\hat{\sigma}^2}, -\frac{s_k\left(\hat{\beta}_k \boldsymbol{e}^T\right)}{\hat{\sigma}^4} \right)^T_{(p+1)\times n},$$

where $\boldsymbol{b}_k$ denotes a vector composed by $p-1$ zeros and 1 in the $kth$ line.

Response variable perturbation scheme: let $\boldsymbol{Y}_\omega = \boldsymbol{Y} + s_y\boldsymbol{\omega}$ represent the perturbed response variable with $s_y$ denoting the scale factor. In this case, the matrix $\boldsymbol{\Delta}$ is given by

$$\boldsymbol{\Delta} = \left( \frac{s_y \boldsymbol{X}^T}{\hat{\sigma}^2}, \frac{s_y \boldsymbol{e}^T}{\hat{\sigma}^4} \right)^T.$$

Error variance perturbation scheme: let $\sigma^2_{w_i} = \frac{\sigma^2}{w_i}$, $i = 1, \ldots, n$, the corresponding matrix $\boldsymbol{\Delta}$ takes the form

$$\boldsymbol{\Delta} = \left( \frac{\boldsymbol{X}^T D(\boldsymbol{e})}{\hat{\sigma}^2}, \frac{\boldsymbol{e}_{sq}^T}{2\hat{\sigma}^4} \right)^T.$$

Cook (1986) shows that

$$\ddot{L} = -\begin{pmatrix} \frac{\boldsymbol{X}^T\boldsymbol{X}}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{pmatrix}.$$

Afterwards, we apply the LIFS algorithm to the data sets described in the Introduction.

## 3 Applications

To show the usefulness of the proposed methodology we apply the LIFS algorithm to four data sets previously analyzed in the literature and to a simulated data set.

First, we consider the rat data which was considered in Cook (1986) and reported in Weisberg (1980). The data set consists of $n = 19$ observations where the response

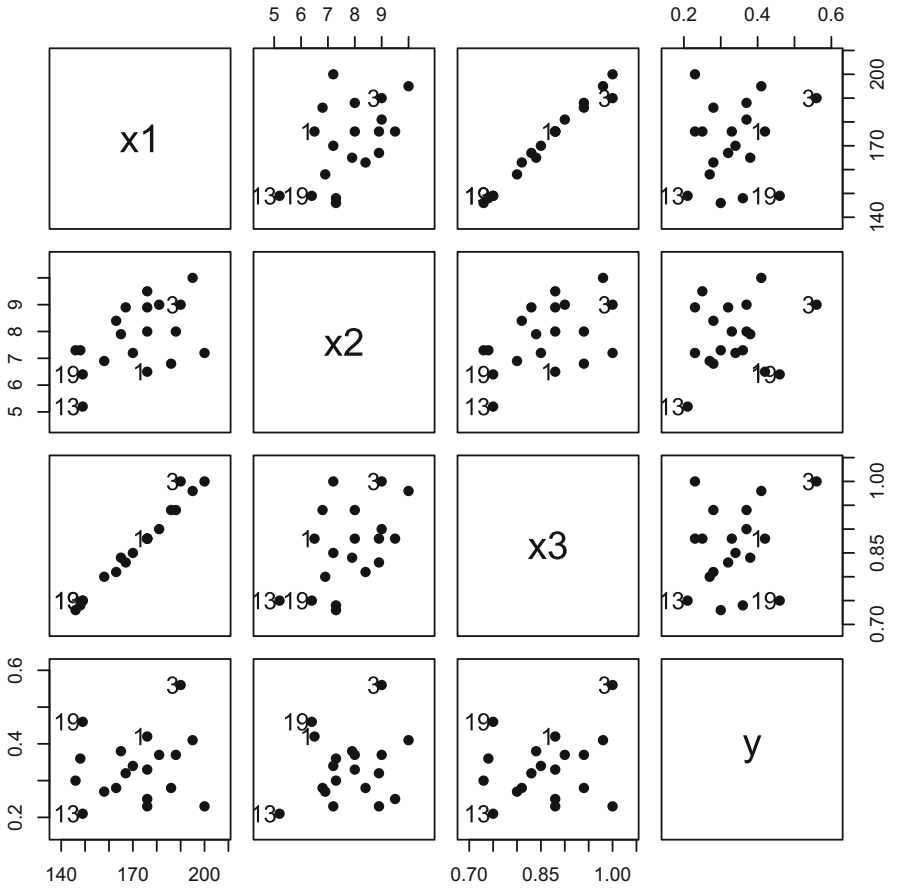


**Fig. 3** Scatter Plot of the rat data variables

variable represents the amount of a particular drug present in the liver of a rat with three covariates: body weight ($X_1$), liver weight ($X_2$) and the relative dose of a drug ($X_3$). The data set was described in the Introduction. Weisberg (1980) concluded that observation 3 is influential and that this rat received a larger dose than it should have received.

Figure 3 shows the scatter plot of the rat data identifying the observations that were detected in Fig. 1.

Cook (1986) considered case-weight perturbation scheme and explanatory variable perturbation scheme. However, it was assumed known variance. Here, we consider that the variance parameter is unknown as was done in Weisberg (1980).

The left hand panel of Fig. 1, presented in the Introduction, shows the forward plot of LIFS algorithm for the rat data considering case-weight perturbation scheme. It starts with a subset of size $m = 5$ and in each step of the algorithm the sample size is incremented by 1, so that in the last step the analysis and the estimation of

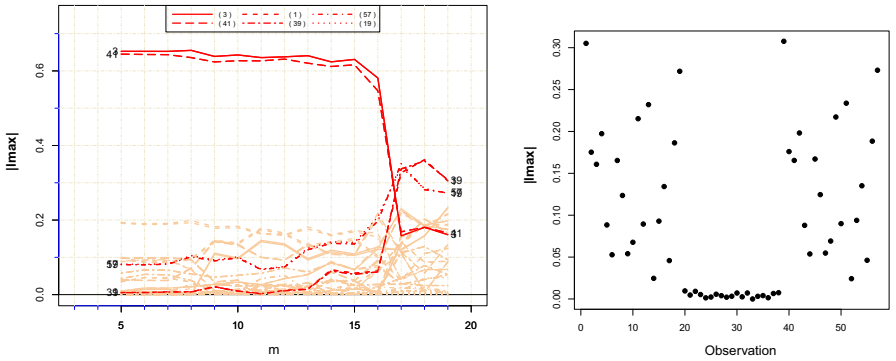


**Fig. 4** Rat data: explanatory variable perturbation scheme. LIFS forward plot (left) and index plot of |**lmax**| (right)

the parameters are done with the whole sample of size $n = 19$, which means that the points in the last iteration are the same as the index plot of |**lmax**| (right hand panel of Fig. 1). Clearly, observation 3 stands out from the rest of observations, as influential, during the evolution of the LIFS algorithm and in the last iteration when observation 3 is introduced into the subset, it becomes masked and observations 19, 13 and 1 pop up. On the other hand, with the usual local influence analysis (right hand panel) it is not possible to identify the observation 3 and the conclusion would be that observations 19, 13 and 1 might be jointly influential.

Next, we perturbed the three covariates using the explanatory variable perturbation scheme. In this case, observations 1 to 19 refer to the elements of the first covariate ($X_1$), while observations 20 to 38 (39 to 57) to the elements of the second covariate ($X_2$) (third covariate ($X_3$)).

The left hand panel of Fig. 4 corresponds to the LIFS forward plot. In this case, observations 3 and 41 stand out from the beginning of the evolution of the LIFS algorithm until the iteration $m = 16$, then these observations are masked, and observations 19 and 57 followed by observations 1 and 39 pop up. In the iteration $m = 17$ observation 3 is introduced in the subset, which makes a steep drop of the observations $(3, 41)$ and a sharp increase of the observations $(1, 39)$ and $(19, 57)$. Observations $(3, 41)$, $(1, 39)$ and $(19, 57)$, are respectively, the $3rd$, $1st$ and $19th$ observation of the covariate $X_1$ and $X_3$, i.e., for instance $(3, 41) = (X_{31}, X_{33})$.

The right hand panel of Fig. 4 shows the index plot of |**lmax**| . In this case, it is clear that covariate $X_1$ and $X_3$ are more influential than the covariate $X_2$.

Table 1 shows the relative changes in the MLE and the respective p-values after dropping observations 3,19, 13 and 1, and also the exclusion of the group of observations (1,3,13,19). With the whole data set, the covariates $X_1$ and $X_3$ are significant at 5% level of significance. After the exclusion of observation 3 or the group of observations (1,3,13,19) none of the covariates is significant.

Figure 5 displays the forward plot of the regression coefficients. On the left hand panel we observe that the value of $\hat{\beta}_3$ changes substantially during the evolution of the forward plot. On the right hand panel $\hat{\beta}_3$ was omitted so that the behavior of the other estimated coefficients during the evolution of the forward plot could be seen. Observe

**Table 1** Parameter estimates, relative changes and the respective p-values

| | Whole | Dropping observations | | | | |
|---|---|---|---|---|---|---|
| | Data set | 3 | 19 | 13 | 1 | (1, 3, 13, 19) |
| $\hat{\beta}_0$ | 0.266 | 0.311 | 0.116 | 0.409 | 0.273 | 0.228 |
| Relative change | | 0.169 | −0.564 | 0.538 | 0.026 | −0.143 |
| p-value | 0.192 | 0.151 | 0.547 | 0.066 | 0.150 | 0.278 |
| $\hat{\beta}_1$ | −0.021 | −0.008 | −0.019 | −0.022 | −0.024 | −0.012 |
| Relative change | | 0.619 | 0.095 | −0.048 | −0.143 | 0.429 |
| p-value | **0.018** | **0.684** | 0.017 | 0.011 | 0.007 | **0.460** |
| $\hat{\beta}_2$ | 0.014 | 0.009 | 0.019 | 0.002 | 0.026 | 0.019 |
| Relative change | | −0.357 | 0.357 | −0.857 | 0.786 | 0.357 |
| p-value | 0.419 | 0.637 | 0.247 | 0.924 | 0.154 | 0.323 |
| $\hat{\beta}_3$ | 4.178 | 1.485 | 3.944 | 4.352 | 4.520 | 2.230 |
| Relative change | | −0.645 | −0.056 | 0.042 | 0.082 | −0.466 |
| p-value | **0.015** | **0.695** | 0.012 | 0.01 | 0.006 | **0.470** |
| $\hat{\sigma}^2$ | 0.005 | 0.005 | 0.004 | 0.004 | 0.004 | 0.003 |
| Relative change | | 0.000 | −0.200 | −0.200 | −0.200 | −0.400 |

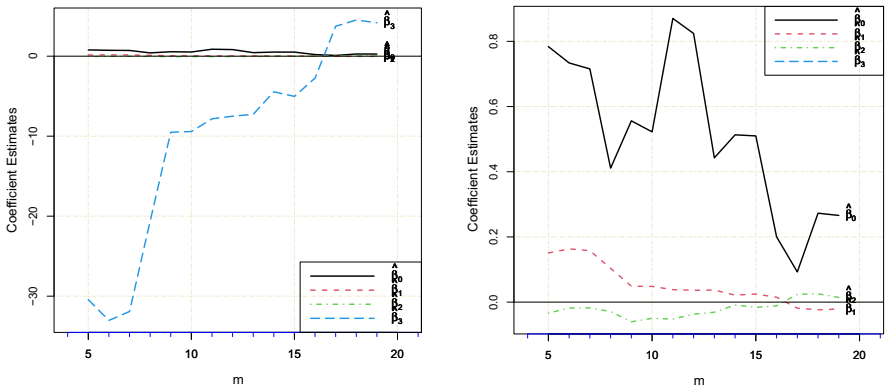


**Fig. 5** Rat data: explanatory variable perturbation scheme. Forward plot of the estimated regression coefficients—all coefficients (left) and without $\hat{\beta}_3$ (right)

that the sign of $\hat{\beta}_1$ ($\hat{\beta}_2$) is positive (negative) during almost the entire evolution of the LIFS algorithm and after iteration $m = 16$ the sign is reversed. The same happens with $\hat{\beta}_3$. The sign of $\hat{\beta}_3$ is negative until iteration $m = 16$ and from iteration $m = 17$ it becomes positive.

The next data set to be considered is the geese data analyzed in Cook (1986) considering that the error variance $\sigma^2$ is known. Here, we also consider $\sigma^2$ as unknown parameter in our analysis. The snow geese data, was conducted to investigate the reliability of the estimate of the number of geese in the flock by an experienced person

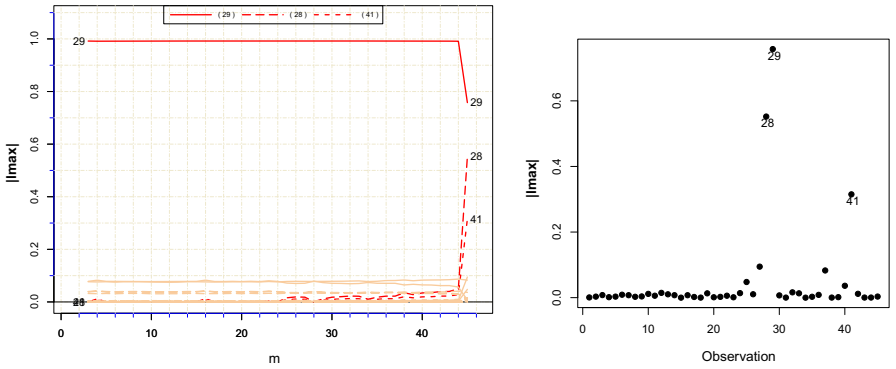


**Fig. 6** Geese data: error variance perturbation scheme. LIFS forward plot (left) and index plot of |**lmax**| (right)

flying over geese summer range areas. The response ($Y$) and explanatory ($X$) variables were described in the Introduction, with $n = 45$ observations.

The LIFS forward plot considering the case-weight perturbation scheme was presented in Fig. 2 (top panel) in the Introduction and it clearly shows that observation 29 followed by the observations 28 and 41 stands out from the rest of observations as possible influential observations during the entire evolution of the LIFS algorithm. Moreover, clearly there is no masked observation. This is what we would expect after looking at the index plot of |**lmax**| (bottom left hand panel of Fig. 2), but it may not be the case as we saw in the previous example (rat data).

Next we consider the variance perturbation scheme. The left hand panel of Fig. 6 shows the LIFS forward plot for the geese data. Clearly, observation 29 appears as the most influential in the entire process of the LIFS algorithm and in the last iteration when observation 29 is introduced in the subset, observation 28 and 41 become influential and the influence of observation 29 is diminished. The right hand panel shows the index plot of |**lmax**|. Observation 29 followed by observations 28 and 41 stand out as possible influential observations.

The left hand panel of Fig. 7 shows the plot of $LD(\boldsymbol{\omega}(a))$ versus $a$ with $\boldsymbol{\omega}(a) = \boldsymbol{\omega}_o + a\boldsymbol{l}$ and $a \in [-1, 1]$ along the directions $\boldsymbol{l} = \boldsymbol{l}_i$, with $i = 28, 29$ and $41$. $\boldsymbol{l}_i$ is the null vector of size 45 with the $ith$ element replaced by 1. Observation 29 is the most influential and observation 41 the least influential among these three observations.

Considering the response variable perturbation scheme (Fig. 8, left hand panel) the LIFS forward plot shows that observation 29 stands out as the most influential observation during the entire evolution of the LIFS algorithm except for the last iteration, when observation 29 is introduced into the subset and observations 28 and 41 pop up. The index plot of |**lmax**| for the last iteration can be seen in the right hand panel and clearly observation 28 stands out as the most influential followed by the observations 29 and 41. On the other hand, the left hand panel shows that observation 29 is the most influential and it was masked in the last iteration.

In addition there are 5 observations that stand out between the observation 29 and the rest of the observations. These are observations 37, 30, 40, 33 and 26 (in this order,

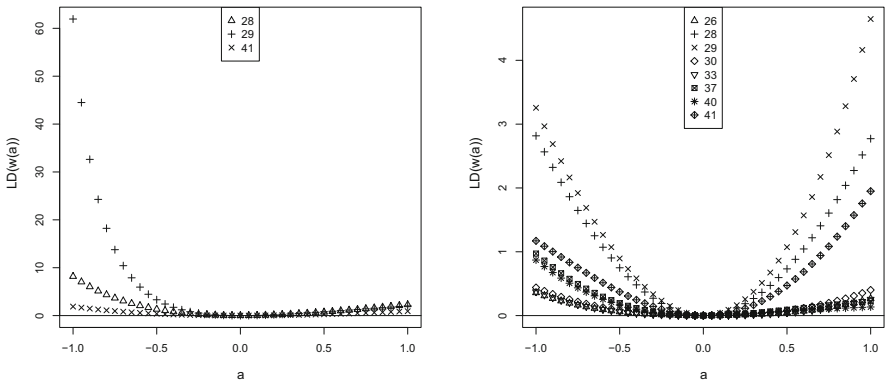


**Fig. 7** Geese data: error variance perturbation scheme (left panel) and response variable perturbation scheme (right panel). Plot of $LD(\boldsymbol{\omega}(a))$ versus $a$

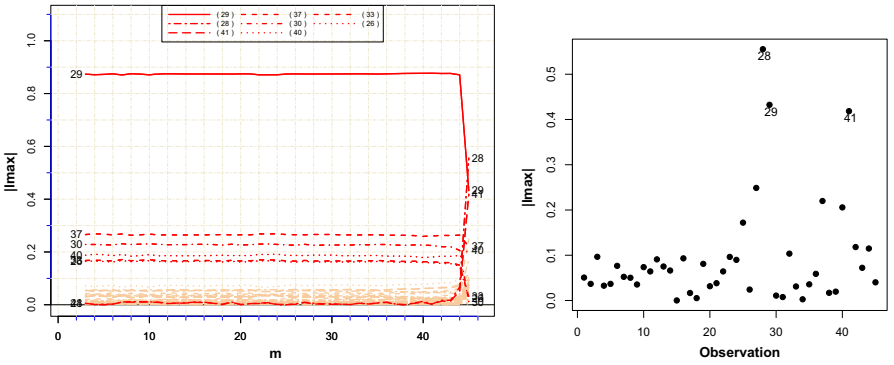


**Fig. 8** Geese data: response variable perturbation scheme. LIFS forward plot (left) and index plot of |**lmax**| (right)

from top to bottom). Observations 30, 33 and 26 were also masked in the last iteration with the entrance of the observation 29.

The right hand panel of Fig. 7 shows the plot of the likelihood displacement along the directions $\boldsymbol{l} = \boldsymbol{l}_i$, with $i = 26, 28, 29, 30, 33, 37, 40$ and 41. Observation 29 is the most influential followed by the observations 28 and 41.

The left hand panel of Fig. 9 shows the evolution of the estimated regression parameters during the LIFS algorithm considering the response variable perturbation scheme. In the last iteration, the estimated value of the intercept changes abruptly when observation 29 is introduced in the subset.

The right hand panel of Fig. 9 shows the evolution of the t-*statistic* for the regression coefficients during the LIFS algorithm. The hypothesis that the intercept is null is rejected in the last iteration with significance level of 5%. Notice that before the last iteration the intercept is significant.

Table 2 shows the MLE of the parameters, relative changes and p-values considering the whole data set and dropping observations 29, 28 and 41, individually and in group.

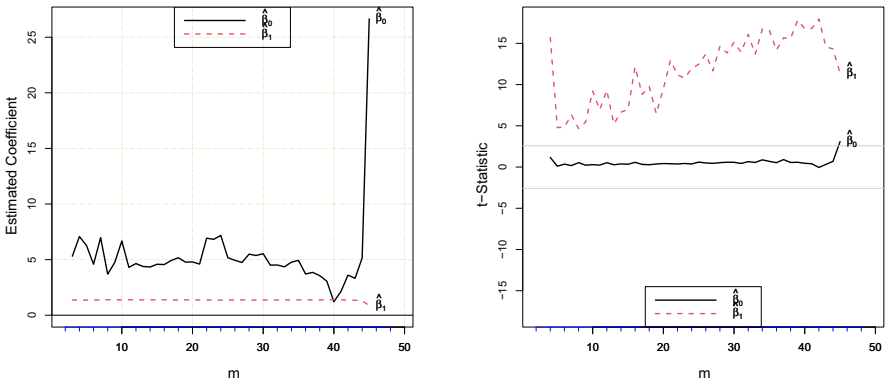


**Fig. 9** Geese data: response variable perturbation scheme. Forward plot of the estimated regression parameters (left) and forward plot of t-*statistic* (right)

**Table 2** Parameter estimates, relative changes and the respective p-values

| | Whole | Dropping observations | | | |
|---|---|---|---|---|---|
| | data set | 29 | 28 | 41 | (28,29,41) |
| $\hat{\beta}_0$ | 26.650 | 5.141 | 29.707 | 27.409 | 18.568 |
| Relative change | | −0.807 | 0.115 | −0.028 | −0.303 |
| p- value | **0.003** | **0.493** | 0 | 0.001 | 0.005 |
| $\hat{\beta}_1$ | 0.883 | 1.280 | 0.782 | 0.831 | 0.964 |
| Relative change | | 0.449 | −0.114 | −0.059 | 0.092 |
| p-value | 0 | 0 | 0 | 0 | 0 |
| $\hat{\sigma}^2$ | 1884.226 | 1058.830 | 1251.536 | 1563.062 | 638.371 |
| Relative change | | −0.438 | −0.336 | −0.170 | −0.661 |

Observe that when observation 29 is removed from the data set the intercept becomes non significant.

The third data set to be analyzed is the bank data which was considered in Riani et al. (2014) and refers to the amount of money made from personal customer over a year. They considered 13 potential explanatory variables (see Riani et al. (2014) for more details) which describe the services used by the customers. The main interest was to discover which activities are particularly profitable. The data set contains 1949 observations.

Considering the variance perturbation scheme, Fig. 10 shows the LIFS forward plot. There is a group of observations (1324, 396, 1338) that stands out during almost the entire evolution of the LIFS algorithm as possibly the most influential observations in the data set, followed by the second group of observations, (86, 892, 1098). Moreover, there are no masked observations. The interesting feature of this plot is that it can also be used as a validation method in this case, as the group of observations stands out as influential observations in almost the entire evolution of the LIFS methodology.

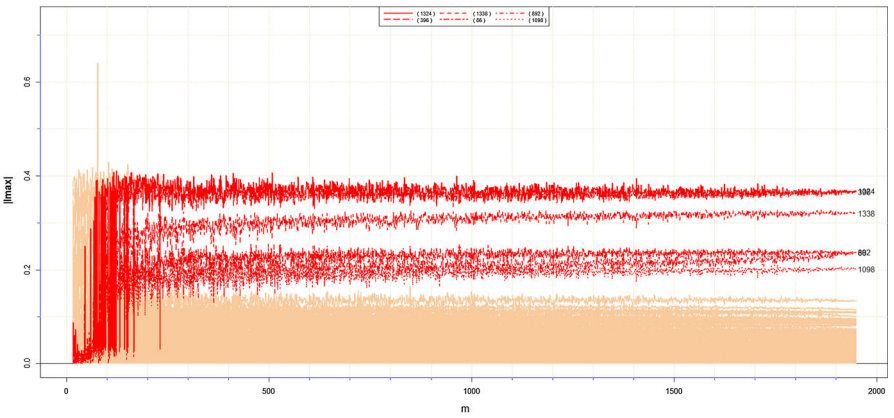


**Fig. 10** Bank data: response variable perturbation scheme. LIFS forward plot

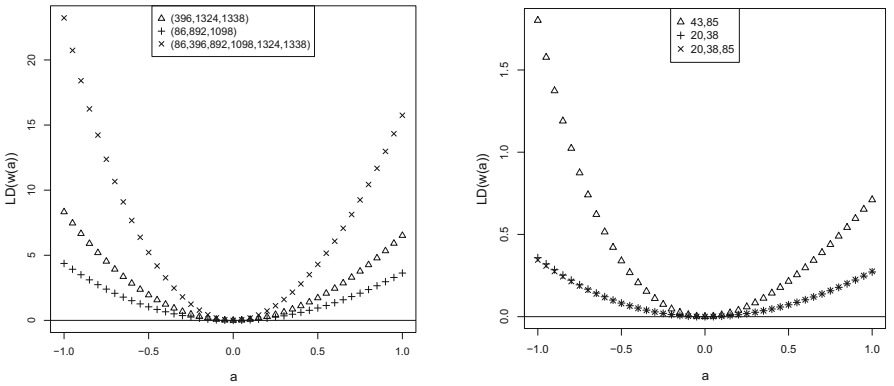


**Fig. 11** Bank data: error variance perturbation scheme (left panel) and customer data: case-weight perturbation scheme (right panel). Plot of $LD(\boldsymbol{\omega}(a))$ versus $a$

The left hand panel of Fig. 11 shows the plot of the likelihood displacement in the direction of the group of observations (1324, 396, 1338), (86, 892, 1098) and also all 6 observations. The group of observations composed by the 6 observations that stands out during almost the entire evolution of the LIFS algorithm, is the most influential group, followed by the first group of observations, (1324, 396, 1338), and the second group of observations, (86, 892, 1098).

In addition, with the whole data set the covariate $X_{10}$ is significant at 1% level, but after the exclusion of the group of observations (1324, 396, 1338) or all the 6 observations, the covariate $X_{10}$ is no longer significant. Considering the parameter estimates, if we drop the first group of observations; the second group of observations or both group of observations, the biggest changes in the parameter estimates occur in the regression coefficients of $X_5$ and $X_{12}$, when the second group is excluded. The relative change in the parameters estimate are 75.36% and 61.40%, respectively. The case weight perturbation scheme reaches the same results and will be omitted.

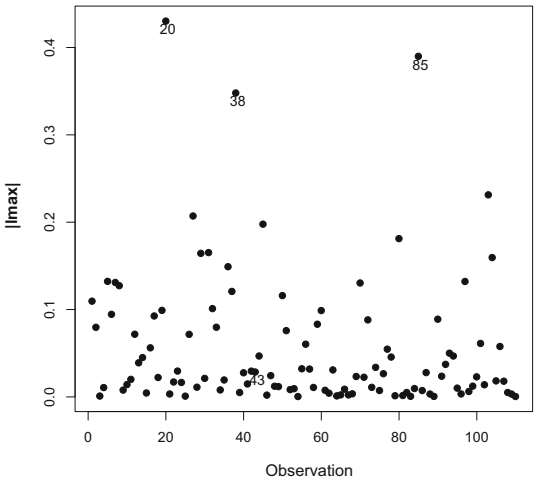


**Fig. 12** Customer data set: case-weight perturbation scheme. Index plot of |**lmax**|

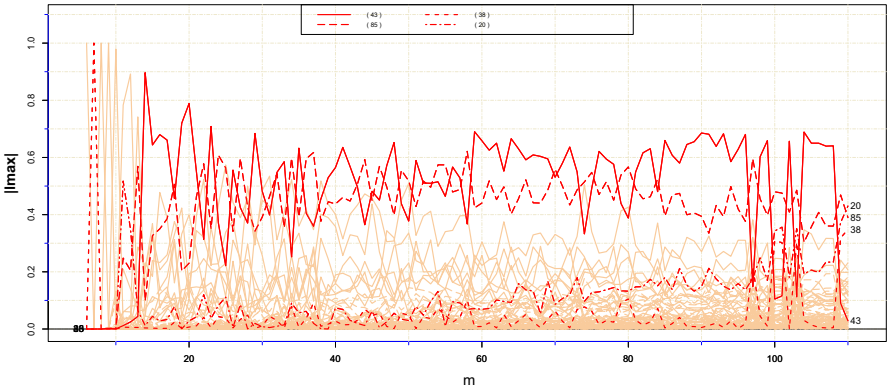


**Fig. 13** Customer data set: case-weight perturbation scheme. LIFS forward plot

The fourth data set to be considered is the customer data set and the corresponding model as reported in Neter et al. (1996). The data set refers to a survey conducted in a two-week period. Initial selection of predictors was conducted which led to the retention of five predictors: $X_1$ representing the number of housing units/1000; $X_2$ the average income/1000, in dollars; $X_3$ the average housing unit age, in years; $X_4$ the distance to nearest competitor, in miles and $X_5$ the distance to store, in miles. The response variable is the number of customers who visited the store from census tract. The Poisson regression model with response function $\mu = exp(\boldsymbol{X}^T\boldsymbol{\beta})$ was fitted to the data. Paula (2013) also analyzed this data set.

The observed information matrix and the matrix $\boldsymbol{\Delta}$ considering the case-weight perturbation scheme, necessary to perform the local influence analysis, can be found in Paula (2013), Sect. 1.10.4. Figure 12 shows the index plot of |**lmax**|. Observations 20, 38 and 85 appears as possible influential observations.

The forward plot of the LIFS algorithm is presented in Fig. 13. Observations 43 and 85 stands out as influential observations during almost the entire evolution of the

**Table 3** Parameter estimates, relative changes, p-values and deviances

| | Whole | Dropping observations | | | |
|---|---|---|---|---|---|
| | Data set | 20 | 38 | 43 | 85 |
| $\hat{\beta}_0$ | 2.942 | 2.959 | 2.920 | 2.989 | 2.921 |
| Relative change | | 0.006 | 0.007 | 0.016 | 0.007 |
| p-value | 0 | 0 | 0 | 0 | 0 |
| $\hat{\beta}_1$ | 0.606 | 0.622 | 0.636 | 0.590 | 0.625 |
| Relative change | | 0.027 | 0.049 | 0.026 | 0.031 |
| p-value | 0 | 0 | 0 | 0 | 0 |
| $\hat{\beta}_2$ | −0.012 | −0.013 | −0.012 | −0.012 | −0.012 |
| Relative change | | 0.073 | 0.014 | 0.001 | 0.005 |
| p-value | 0 | 0 | 0 | 0 | 0 |
| $\hat{\beta}_3$ | −0.004 | −0.005 | −0.003 | −0.004 | −0.003 |
| Relative change | | 0.211 | 0.106 | 0.187 | 0.126 |
| p-value | **0.037** | 0.014 | **0.063** | 0.015 | **0.07** |
| $\hat{\beta}_4$ | 0.168 | 0.172 | 0.165 | 0.174 | 0.173 |
| Relative change | | 0.023 | 0.021 | 0.036 | 0.027 |
| p-value | 0 | 0 | 0 | 0 | 0 |
| $\hat{\beta}_5$ | −0.129 | −0.126 | −0.131 | −0.134 | −0.131 |
| Relative change | | 0.020 | 0.014 | 0.039 | 0.014 |
| p-value | 0 | 0 | 0 | 0 | 0 |
| deviance | 114.985 | 111.242 | 110.988 | 111.978 | 112.116 |
| Relative change | | 0.033 | 0.035 | 0.026 | 0.025 |

LIFS algorithm. In the final steps, observation 43 is masked and observations 20 and 38 pop up.

The right hand panel of Fig. 11 shows the plot of the likelihood displacement in the direction of the group of observations (43, 85), the two observations that stand out during almost the entire evolution of the LIFS algorithm, observations (20, 38), the two observations that pop up in the final steps and observations (20, 38, 85), the three observations that stands out in the index plot of $|\mathbf{lmax}|$. The plot of $LD(\boldsymbol{\omega}(a))$ versus $a$ shows that if we perturb in the direction of the group of observations (43, 85) the likelihood displacement increases more than if we perturb in the direction of the group of observations (20, 38), whereas the joint influence of the group of the observations (20, 38) are the same as the group of observations (20, 38, 85).

Considering the global influence, the exclusion of observations 38 or 85 change the inferential results. Table 3 shows that the covariate $X_3$ is not significant at 5% level of significance, however after dropping the observation 38 or 85 it becomes significant.

The last data set to be analyzed is a simulated data set where nearly 50% of the observations are contaminated. Considering the simple linear regression model it was generated 76 observations according to Fig. 14, where the 18 observations in green and the 18 observations in red represent the contaminated portion of the data set and the majority (40 observations) are in beige.

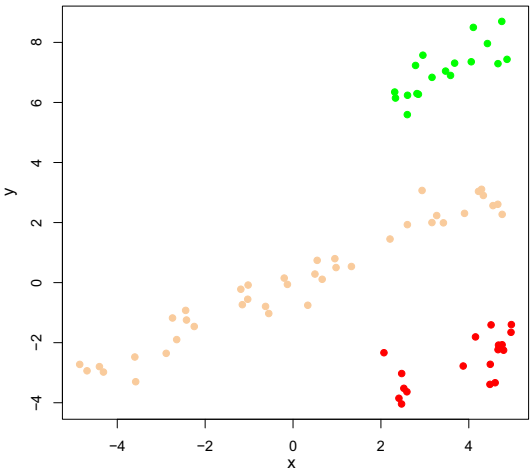


**Fig. 14** Scatter Plot of the simulated data

The LIFS methodology was applied to the simulated data set considering linear regression model with the use of the explanatory variable perturbation scheme and the response variable perturbation scheme. Figure 15 top (bottom) presents the LIFS forward plot for the explanatory variable perturbation scheme (response variable perturbation scheme). Both plots show that observations in red and green may be influential, while the observations in beige are robust to the induced perturbations.

## 4 Conclusions

Influence diagnostics is an important step in statistical data analysis. If there are observations that can influence the results of an analysis, these observations should be known. Thought there are many methodologies to find out influential observations, these methodologies may fail to detect masked influential observations.

Atkinson and Riani (2000) proposed the forward search algorithm considering the least-squares estimates and the corresponding residuals to identify masked outliers, in regression models, considering the global influence. On the other hand, Cook (1986) developed a general methodology to identify multiple influential observations considering the local influence. However, this methodology may not detect masked locally influential observations. To fill this gap, we proposed the LIFS algorithm to detect masked influential observations considering regression models and local influence approach, thought the proposed methodology may be applied to any model with a well-behaved likelihood.

We applied the methodology to four data sets which were previously analyzed in the literature and to a simulated data to detect outliers and/or influential observations. The local influence analysis were performed and influential observations were obtained, however, as can be seen in Sect. 3 there are masked influential observations that were not detected with the usual local influence analysis. The proposed methodology successfully detected masked influential observations and the applications also revealed

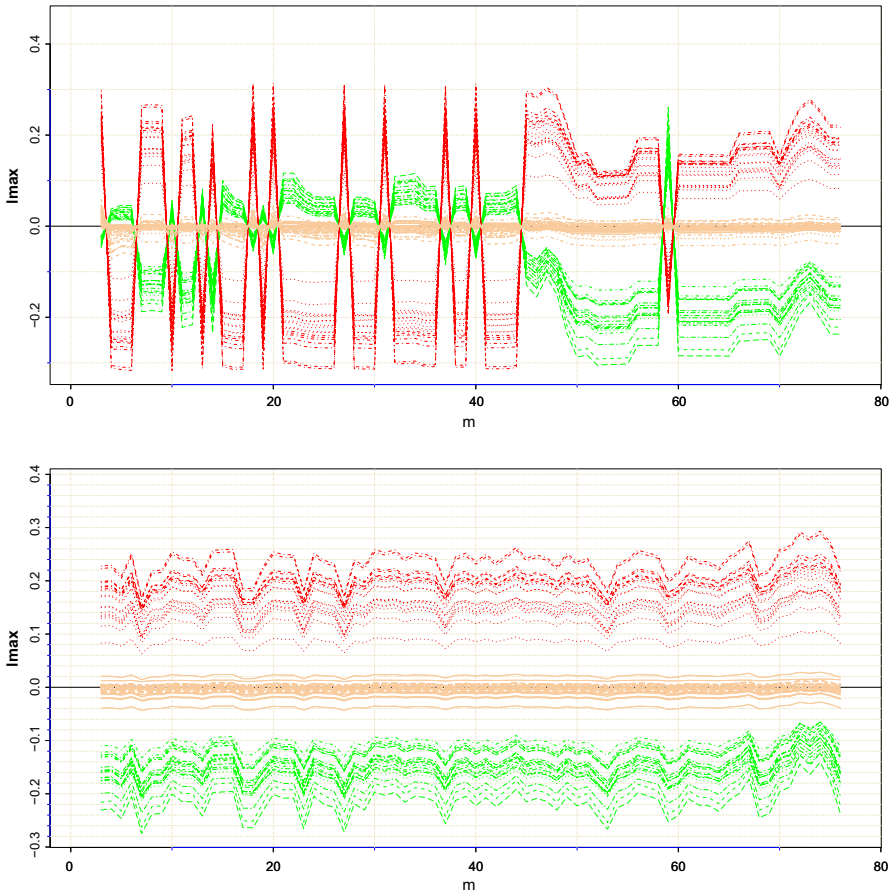


**Fig. 15** Simulated data: LIFS forward plot, explanatory variable perturbation scheme (top) and response variable perturbation scheme (bottom)

some interesting aspects of the LIFS algorithm in the local influence analysis. Figures 5 and 9 , for instance, show how the quantities of interest evolve during the iterations of the LIFS algorithm and even if there are no masked influential observations it can give extra information, as validation (see Fig. 2).

Hence the LIFS algorithm may be used to complement the local influence analysis proposed by Cook (1986) to detect masked influential observations.

**Acknowledgements** We would like to gratefully thank the anonymous referees for their comments and suggestions that definitely helped to improve the quality of the paper. The research was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References


Atkinson A, Riani M (2000) Robust diagnostic regression analysis. Springer, New York

Atkinson AC, Riani M, Cerioli A (2018) Cluster detection and clustering with random start forward searches. J Appl Stat 45:777–798

Beckman RJ, Nachtsheim CJ, Cook RD (1987) Diagnostics for mixed-model analysis of variance. Technometrics 29(4):413–426

Belsley DA, Kuh JA, Edwin K, Welsch JA, Roy E (1980) Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New York

Cerioli A, Farcomeni A, Riani M (2019) Wild adaptive trimming for robust estimation and cluster analysis. Scand J Stat 46(1):235–256

Cerioli A, Riani M, Atkinson AC, Corbellini A (2018) The power of monitoring: how to make the most of a contaminated multivariate sample. Stat Methods Appl 27(4):559–587

Chatterjee S, Hadi AS (1988) Sensitivity analysis in linear regression, Volume XIV of Wiley series in probability and statistics

Cook RD (1977) Detection of influential observation in linear regression. Technometrics 19(1):15–18

Cook RD (1986) Assessment of local influence. J R Stat Soc Series B 48(2):133–169

Cook RD, Weisberg S (1982) Residuals and influence in regression. Monographs on statistics and applied probability

Escobar LA, Meeker WQ Jr (1992) Assessing influence in regression analysis with censored data. Biometrics 48:507–528

Galeano P, Peña D (2019) Data science, big data and statistics. TEST 28(2):289–329

Grané A, Manzi G, Salini S (2021) Smart visualization of mixed data. Stats 4(29):472–485

Hadi AS (1992) Identifying multiple outliers in multivariate data. J R Stat Soc Ser B 54:761–771

Hadi AS, Simonoff JS (1993) Procedures for the identification of multiple outliers in linear models. J Am Stat Assoc 88(424):1264–1272

Labra FV, Aoki R, Rojas F (2007) An application of the local influence diagnostics to ridge regression under elliptical model. Commun Stat Theory Methods 36(4):767–779

Lawrance AJ (1988) Regression transformation diagnostics using local influence. J Am Stat Assoc 83(404):1067–1072

Mavridis D, Moustaki I (2009) The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. J Comput Graph Stat 18:1016–1034

Montgomery D, Jennings CL, Kulahci M (2011) Introduction to time series analysis and forecasting, vol 526. Wiley, Hoboken

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models. Irwin Professional Publishing, Burr Ridge

Paula GA (1993) Assessing local influence in restricted regression models. Comput Stat Data Anal 16(1):63–79

Paula GA (2013) Modelos de regressão: com apoio computacional. IME-USP São Paulo

Riani M, Atkinson AC, Cerioli A, Corbellini A (2019) Comments on: data science, big data and statistics. TEST 28:349–352

Riani M, Cerioli A, Atkinson AC, Perrotta D et al (2014) Monitoring robust regression. Electron J Stat 8(1):646–677

Russo CM, Paula GA, Aoki R (2009) Influence diagnostics in nonlinear mixed-effects elliptical models. Comput Stat Data Anal 53(12):4143–4156

Russo CM, Paula GA, Cysneiros FJA, Aoki R (2012) Influence diagnostics in heteroscedastic and/or autoregressive nonlinear elliptical models for correlated data. J Appl Stat 39(5):1049–1067

Thomas W, Cook RD (1990) Assessing influence on predictions from generalized linear models. Technometrics 32(1):59–65

Weisberg S (1980) Applied linear regression, 1st edn. Wiley, Hoboken

Zhu F, Liu S, Shi L (2016) Local influence analysis for poisson autoregression with an application to stock transaction data. Statistica Neerlandica 70(1):4–25