A CLASSIFICATION OF SINGLE INFLUENTIAL OBSERVATION STATISTICS IN REGRESSION ANALYSIS

Assoc. Prof. Dr. Suat SAHİNLER
Mustafa Kemal University, Agriculture Faculty, Biometry and Genetics Unit, Hatay/TURKEY

Prof. Dr. Yüksel BEK
19th May University, School of Medicine, Department of Biostatistics, Samsun / TURKEY

1. Introduction

The results of least squares fit of the general linear model

$$Y=X\beta+\varepsilon \quad (1)$$

to given data set can be substantially influenced by omission or addition one or few observations. Therefore, the least squares method does not ensure that the regression model proposed is fully acceptable from the statistical and physical points of view. Usually one of the main problems is that all observations have not an equal influence in least squares fit and in the conclusions that result from such analysis.

The detection, assessment, and understanding of influential observations are the major areas of interest in regression model building. It is important for a data analyst to be able to identify influential observations, assess and understanding their effects on various aspects of the analysis. They are rapidly gaining recognition and acceptance by practitioners as supplements to the traditional analysis of residuals. Residuals play an important role in regression diagnostics; no analysis is complete without a thorough examination of the residuals. The standard analysis of regression results is based on certain assumptions (for more information we refer to [1], [2]). It is necessary to check the validity of these assumptions before drawing conclusions from an analysis [3].

Definitions

An observation may not have the same influence on all regression results. Therefore, the observations which are used in the regression analysis can be examined under four groups.

a) Usual observations: It is considered that the observations which have equal effects to the important properties as fitted values, estimated parameters of the regression analysis can be called usual observations.

b) Outliers: An outlier is an observation for which the studentized residual t_i^* is large in magnitude compared to other observations in the data set. These observations

may indicate violation of assumptions and perhaps the need for an alternative model.

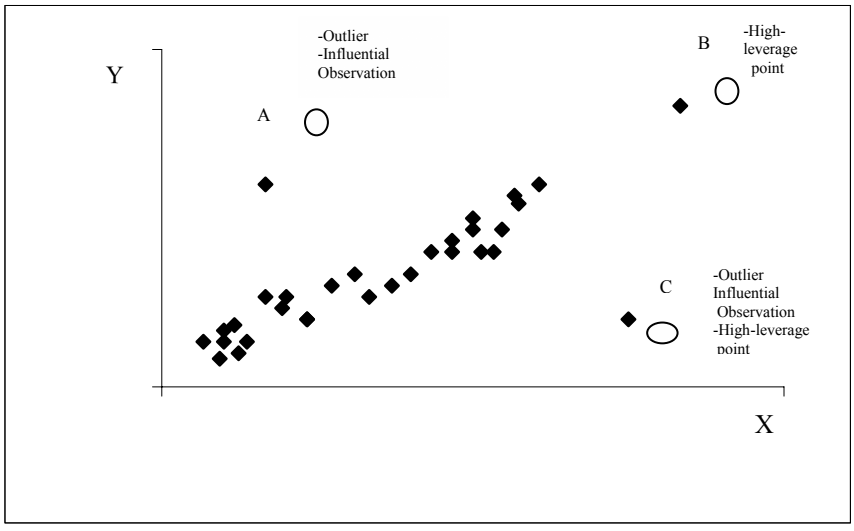
c) High-Leverage Points: If only the space of X is considered, a high leverage point is an observation far from the center of X space compared to the other observations [4], [5], [6]. Observations far from the center of the predictors have low variance residuals ($Var(e_i)=(1-h_{ii})\sigma^2$). This reflects the fact that such observations have high-leverage, that is, that they pull the regression line towards themselves. Points with high leverage may be regarded as outliers in the X space. The concept of leverage is linked entirely to the predictor variables and not to the response variable.

d) Influential observations: Influential observations are those observations that, excessively influence the fitted regression equation, regression coefficients and the estimates of variance (σ^2) compared to other observations in the data set [6].

Suppose we have the data points as in Figure 1 and refer to three points marked by letters A, B and C.If point A is considered in Figure 1, it will not be a high-leverage point because it is close to the center of X, but it will be clearly be an outlier and an influential point. It will have a large residual, and its inclusion may not change the slope but will chance the intercept of the fitted line. Its inclusion will also change the estimated error variance, and hence the variances of the estimated coefficients.

Figure 1: An example illustrating the distinction among outliers, high-leverage points and influential observations.

If point B is considered for inclusion, it will have a small residual because its Y position is near where the line passes through other points. It will be a high-leverage point because it is an outlier in X. However, it will not have a large influence on the fitted regression equation. It is clear that point B en example of a high-leverage point



that is neither an outlier nor an influential point. Note also that point B is an extreme

point in both X and Y, yet it is not influence on the estimated regression coefficients (because point B is an extreme point in X space, it may however be influential on the standard error of the regression coefficients)

If we consider point C, it was seen that C would be an outlier, a high-leverage point and an influential point. It will be an outlier because it will have a large residual. It will be a high-leverage point because it is an extreme point in X space. It is an influential observation because its inclusion will substantially chance the characteristics of the fitted regression equation.

We can also note the following remarks from these definitions and Figure 1.

-Outliers need not be influential observations.

-Influential observations need not be outliers.

-There is a general tendency for high-leverage points to have small residuals and to influence the fit disproportionately.

-High-leverage points need not be influential observations and influential observations are not necessarily high-leverage points. However, high-leverage points are likely to be influential.

-An observation might be an outlier, leverage point or influential observation simultaneously.

These examples point up the fact that, examination of residuals alone may not detect aberrant or unusual observations such as those indicated by B in Figure 1. Informal graphical methods or formal testing procedures based on the residuals will fail to detect these unusual points. Observations with these characteristics which have small residuals and highly influential on the fit often occur in real-life data. Statistical measures for assessing leverage and influence are, therefore, needed.

This paper provides a short survey of single points influence diagnostics, illustrated with the real data sample consisting of ages (Y) and otolith length (X_l) of fish which were taken in fisheries faculty from 138 fish respectively. It was changed the 40 *th*, 80 *th*, and 120 *th* observations in data set as A, B and C points in Figure 1 to show how they characterize the influence of these cases in data and to test the sensitivity of the statistics. Hence the regression equation was fitted, the statistics were calculated, the influential observations were identified, and the results concluded for the real data set.

2. Diagnostics for Identifying Unusual Observations

It was reminded some statistics briefly for measuring the effects of a point on some regression results following.

2.1. Examination of residuals

Because of having non-constant variance and being not often indicate strongly deviant points, the ordinary residuals are not appropriate for diagnostic purpose. Therefore, a transformed version of them is preferable [7]. These are the normalized

residual, the standardized residual, the internally studentized residual and the externally studentized residual (jackknife residuals) and calculated as respectively,

The normalized and standardized residuals are defined as respectively;

$$e_{in} = \frac{e_i}{\sqrt{(e'e)}} \tag{2}$$

$$e_{is} = \frac{e_i}{\sqrt{(e'e)/(n-p)}} \tag{3}$$

It is falsely assumed that these residuals are normally distributed quantities with zero mean and variance equal to one, but in reality these residuals have non-constant variance. When these residuals are used for identifying the outliers, $\pm 3\sigma$ is classically recommended as a calibration point, but this approach is quite misleading, and may cause wrong decisions to be taken regarding data [7].

The internally studentized residual

$$t_i = \frac{e_i}{\sqrt{(e'e)/(n-p)}\sqrt{1-h_{ii}}} \tag{4}$$

where $h_{ii} = x'_i(X'X)^{-1} x_i$, i = 1,2,..., n. The internally studentized residuals behave much like a Student's t (t_{n-p}) random variable except for the fact that the numerator and denominator of t_i are not independent.

The externally studentized residual

$$t_i^* = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}} \tag{5}$$

where , $s^2_{(i)} = [s^2(n-p)-e_i^2/(1- h_{ii})]/(n-p-1)$ is the residual mean squared error estimate of σ^2 comes from refitting model without observation i and is robust to problems of gross errors in the i *th* observation. t_i^* preferred over other transformed residuals [8], [9].

-Because the t_i^* exactly follows a t_{n-p-1} distribution for which tables are readily available under the normality assumption, it can be simply assessed the magnitude of to determine if point i is an outlier.

-t_i^* reflects large deviations more dramatically than do the others.

2.2. The Diagonal Elements of H = X(X'X) $^{-1}$ X' Matrix

The i *th* diagonal elements of H matrix is defined as,

$$h_{ii} = x'_i(X'X)^{-1} x_i \text{ , } i = 1,2,...,n \quad (6)$$

and used for identifying high-leverage points. Leverage is the potential for an observation to affect the fit of the model. h_{ii} is a standardized distance of the i *th* observation to $\bar{x}$. Large h_{ii} means the observation is far from $\bar{x}$, small h_{ii} means it is near the center of the predictors [9], [10]. Thus, h_{ii} represents the high-leverage of the i *th* observation y_i in determining its own predicted value $\hat{y}_i$. In fact, in balanced design $\sum h_{ii} = tr(H) = p$, so the average value of the h_{ii}'s is p/n. This is a rough rule of thumb and some calibration points are suggested for various p and n-pin Table 1 [6], [11]. h_{ii} values grater than calibration points in Table 1 are cause of concern.2.3. Mahalanobis Distance (MD_i)

Mahalanobis distance is defined as:

$$MD_i = \sqrt{(x_i - \mu_x)'\Sigma^{-1}(x_i - \mu_x)}, i = 1,2,...,n \quad (7)$$

where μ_x is the mean of X and Σ^{-1} is the inverse variance-covariance matrix of X and used as a measure of the leverage of an observation. Mahalanobis distance weights the distance of a data point x_i from its mean μ_x such that observations that are on the same multivariate normal density contour will have the same distance [9]. Mahalanobis distances are approximated by the $\chi^2_{1-\alpha,p}$ distribution, where p is the number of parameters [12], [13].

2.4. Weighted Squared Standardized Distance ($WSSD_i$)

Weighted sum of squared distance is a measure of the sum of squared distance of x_{ij} from the average of the j *th* variable, X_j, weighted by the relative importance of the j *th* variable and defined as,

$$WSSD_i = \sum_{j=1}^{k} c_{ij}^2 / s_Y^2 \text{ , i=1,2,…, n} \quad (8)$$

where $c_{ij} = \hat{\beta}_j(x_{ij} - \bar{x}_j)$, i=1,2,…,n, j=1,2,…,k, where $\bar{x}_j$ is the average of the j *th* column of X and used as a measure of the leverage of an observation. In simple regression case $WSSD_i$ is equivalent h_{ii} and for this reason, the calibration point for h_{ii} in Table 1 can be used for $WSSD_i$ [6].

	Condition		Calibration point
If	p<6 and (n-p)>12	Then	3p/n
	2<p<6 and (n-p)>30		2.5p/n
	6≤ p<15 and (n-p)>30		2p/n
	p>15 and (n-p)>30		1.5p/n

Table 1: The calibration points of h_{ii}, and $WSSD_i$ statistics for various p and n-p.

2.5. Andrews-Pregibon Statistic (AP_i)

The Andrews-Pregibon Statistic is based on the volume of the confidence ellipsoids and calculated as

$$AP_i = 1- h_{ii} - (e_i^2 /e'e) = 1 - h_{ii}^* \quad (9)$$

where $h_{ii}^* = h_{ii} + (e_i^2 /e'e)$. AP_i measures the influence of the i *th* observation on the estimated parameters by combining the residual sum of squares and the volume of the confidence ellipsoids when the i *th* observation is omitted. It is used 1-(2(p+1)/n) as a calibration point for AP_i statistic.

2.6. Cook Statistic (C_i)

This statistic is proposed by [14] and calculated as

$$C_i = (t_i^2 /p)(h_{ii} /(1- h_{ii})) \quad (10)$$

where t_i is the internally studentized residual in Equation 4. C_i measures the overall influence of each observation on the regression coefficients, including the intercept. The usual criterion is that a point is influential if exceeds the median of the $F_{p,n-p}$ distribution [15] or $1/F_{n-p,p}$.

2.7. Likelihood Distance (LDi)

Likelihood Distance Statistic for the influence of the i *th* observation on only $\hat{\beta}$, s^2 and ($\hat{\beta}$, s^2) simultaneously are measured by the following equations

$$LD_i(\beta) = n \log[pC_i /(n-p)+1] = n \log[b_i h_{ii} /(1-h_{ii})+1] \quad (11)$$

$$LD_i(\sigma^2)= n \log[n/(n-1)]+n \log(1-b_i)+b_i(n-1)/(1-b_i)-1 \quad (12)$$

$$LD_i(\beta,\sigma^2)=n \log[n/(n-1)]+n \log(1-b_i)+b_i(n-1)/[(1-b_i)(1-h_{ii})]-1 \quad (13)$$

where $b_i= t_i^2/(n-p)$. $LD_i(\beta)$ and $LD_i(\sigma^2)$ use $\chi^2_{1-\alpha,p}$ and $LD_i(\beta,\sigma^2)$ uses $\chi^2_{1-\alpha,p+1}$ as a calibration point.

2.8. Covariance Ratio Statistic (CVR_i)

CVR_i examines how the precision of the parameter estimates change with the removal of the i *th* observation. The CVR_i measures the change in $|(\text{var}(\hat{\beta})|$ and considers the ratio of $\det(s^2_{(i)}(X'_{(i)}X_{(i)})^{-1})$ to $\det(s^2(X'X)^{-1})$:

$$CVR_i = \det\{s^2_{(i)}(X'_{(i)}X_{(i)})^{-1}\} / \det\{s^2(X'X)^{-1}\}$$

$$= (s^2_{(i)}/s^2)^p/(1-h_{ii}) = \{(n-p-t_i^2)/(n-p-1)\}^p/(1-h_{ii}) \quad (14)$$

When all observations have equal influence on the covariance matrix, CVR_i is approximately equal to one. If $|CVR_i-1| \geq 3p/n$ than the i^{th} observation is influential on parameter estimates and estimated variance of regression coefficients.

2.9. Welsch-Kuh Statistic (WK_i)

This statistic is calculated as

$$WK_i = |t_i^*|(h_{ii}/(1-h_{ii}))^{1/2} \quad (15)$$

and measures how many standard errors $\hat{y}_i$ moves when the the i *th* observation is omitted [9]. It was called as $DFFITS_i$ by [16] and if $WK_i > 2(p/n)^{1/2}$ than the i *th* observation is influential on $\hat{y}_i$.

2.10. Cook-Weisberg Statistic (CW_i)

This statistic is proposed by [17] as

$$CW_i = (-1/2)\log(CVR_i)+(p/2)\log[(F_{(\alpha,p,n-p)})/(F_{(\alpha,p,n-p-1)})] \quad (16)$$

and measures the influence of the i *th* observation on the volume of confidence ellipsoid for β. If this quantity is large and positive, then deletion of i *th* observation will result in a substantial decrease in volume of confidence ellipsoid and if it is large and negative, the case will result in a substantial increase in volume of confidence ellipsoid.

2.11. Welsch Statistic (W_i)

Welsch statistic has suggested by [18] as

$$W_i = WK_i\,[(n-1)/(1-h_{ii})]^{1/2} \quad (17)$$

and measures the influence of the i *th* observation on both s^2 and estimation of regression coefficient β. Welsch statistic uses $3p^{1/2}$ if n>15 as a calibration point and gives more emphasis to high-leverage points.

2.12. Modified Cook Statistic (C_i^*)

Cook statistic has modified by replacing s^2 by $s^2_{(i)}$, taking the square root of C_i and adjusting C_i for the sample size. Thus,

$$C_i^* = WK_i\,[(n-p)/p)]^{1/2} \tag{18}$$

and this modification improves C_i in following ways;

- C_i^* gives more emphasis to extreme points,

- C_i^* becomes more suitable and identical for graphical displays.

C_i^* statistic uses $2[(n-p)/n]^{1/2}$ as a calibration point and measures the influence of the i *th* observation on both s^2 and estimation of regression coefficient β.

2.13. $DFBETAS_{j,i}$ Statistic

$DFBETAS_{j,i}$ is calculated as

$$DFBETAS_{j,i} = (\hat{\beta}_j - \hat{\beta}_{j(i)})/(s_{(i)}\sqrt{C_{jj}}) \tag{19}$$

where C_{jj} is the j *th* diagonal elements of matrix $C=(X'X)^{-1}$, $\hat{\beta}_j$ and $\hat{\beta}_{j(i)}$ are the estimates of β_j obtained from the full data and the data without the i *th* observation, respectively. This statistic measures the influence of the i *th* observation on the estimation of j *th* regression coefficient β_j. This means that $DFBETAS_{j,i}$ measures how many standard errors β_j moves when the i *th* observation is omitted [9]. C_i is roughly the average of the squares of the $DFBETAS_{j,i}$. If $|DFBETAS_{j,i}|>2n^{1/2}$ than the i *th* observation is influential on the j *th* regression coefficient β_j.

The calibration points, some common properties and differences of the statistics given above were summarized in Table 2.

Used statistics and formula	Calibration point	Common propertiesand differences of the statistics
$e_{in} = \dfrac{e_i}{\sqrt{(e'e)}}$	3σ	1) Each is a transformed version of the ordinary residuals. 2) They identify outliers. 3) Using e_{in} and e_{is} may cause wrong decisions to be taken regarding data. Because, they don't have constant
$e_{is} = \dfrac{e_i}{\sqrt{(e'e)/(n-p)}}$	3σ	

$t_i = \frac{e_i}{\sqrt{(e'e)/(n-p)}\sqrt{1-h_{ii}}}$	t_{n-p}	variance. 4) t_i*is monotonic transformation of t_i , $t_i^* = t_i\sqrt{\frac{n-p-1}{n-p-t_i^2}}$
$t_i^* = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$	t_{n-p-1}	5)if rank($X_{(i)}$)=p, and $\varepsilon \sim N_n(0,\sigma^2)$, then t_i* is distributed as t_{n-p-1}. and reflects large deviations more dramatically than the others. 6) $t_i \approx\approx t_{n-p}$ [9].
$h_{ii} = x'_i(X'X)^{-1} x_i$	Table 1	1) They identify high-leverage points. 2) h_{ii} ignores the information contained in Y. 3) $hii = \frac{MD_i^2}{n-1} + \frac{1}{n}$ 4) $\sum h_{ii} = tr(H) = p$ 5) In simple regression case, $WSSD_i$ is equivalent to h_{ii}.
$MD_i = \sqrt{(x_i - \mu_x)'\Sigma^{-1}(x_i -}$	Table 1	
$WSSD_i = \sum_{j=1}^{k} c_{ij}^2 / s_Y^2$	Table 1	
$LD_i(\sigma^2) = n\log[n/(n-1)] + n\log[1-b_i] + b_i(n-1)/(1-b_i) - 1$	$\chi^2_{1-\alpha,p}$	1) LD_i (σ^2) is based on maximum likelihood function. 2) It uses s as the estimate of σ instead of $s_{(i)}$. 3) It is only one statistic which measures the influence of an observation on variance(σ^2) only. 4) Its values are influenced by outliers but not influenced by high leverage points.
$WK_i = t_i^*(h_{ii}/(1-h_{ii}))^{1/2}$	$2(p/n)^{1/2}$	1)uses $s_{(i)}$ as the estimate of σ instead of s. 2) It is only one statistic which measures the influence of an observation on predicted values
$DFBETAS_{j,i} = (t_i^* C_{ij})/\{(1-h_{ii})C'_j C_j\}^{1/2}$	$2/n^{1/2}$	1) $DFBETAS_{j,i}$ statistic uses $s_{(i)}$ as the estimate of σ instead of s. 2) Its values are influenced by outliers and high leverage points. 3)It measures influence of i *th* Observation on β_i

Table 2: The summaries of the statistics for identifying outliers, high-leverage points and influential observations.

Used statistics and formula	Calibration point	Common propertiesand differences of the statistics
$LD_i(\beta) = n\log[b_i h_{ii}/(1-h_{ii})+1] = n\log[pC_i/(n-p)+1]$	$\chi^2_{1-\alpha,p}$	1)They use s as the estimate of σ instead of $s_{(i)}$ which is robust to problems of gross errors in the i *th* observation. 2) They measure the influence of an observation on regression coefficients (β) only. 3) They are effective in the detection of observations that have influenced on the parameter estimates. 4) LD_i (β) is based on maximum likelihood function. 5) C_i is based on confidence ellipsoids and when n is large, C_i has smaller values. 6) CVR_i statistic is more sensitive to the high leverage point and outliers then the other statistics thereby reduces the ability of CVR_i to detect influential observations. However, it is the most suggested statistics in this group because of identifying influential observations successfully in the data set [4], [19].
$CVR_i = \{(n-p-t_i^2)/(n-p-1)\}^p/(1-h_{ii})$	$\lvert CVR-1\rvert \geq 3p/n$	
$CW_i = -\frac{1}{2}\log(CVR_i) + \frac{p}{2}\log\frac{F_{\alpha,p,n-p}}{F_{\alpha,p,n-p-1}}$	--	
$C_i = (t_i^2/p)(h_{ii}/(1-h_{ii}))$	$1/F_{\alpha,n-p,p}$	
$AP_i = 1-h_{ii}-(e_i^2/e'e) = 1-h_{ii}^*$	$1-[2(p+1)/n]$	1) W_i and C_i^* statistics use $s_{(i)}$ as the estimate of σ instead of s. 2) They measure the influence of an observation on both regression coefficients(β) and variance(σ^2) simultaneously. 3) Their values are influenced by outliers and high leverage points. 4) $LD_i(\beta, \sigma^2)$ is based on maximum likelihood function and if $h_{ii}=0$ then
$LD_i(\beta,\sigma^2) = n\log[n/(n-1)] + n\log(1-b_i) + b_i(n-1)/[(1-b_i)(1-h_{ii})]-1$	$\chi^2_{1-\alpha,p+1}$	
$W_i = WK_i[(n-1)/(1-h_{ii})]^{1/2}$	$3p^{1/2}$	

$C_i^* = WK_i\,[(n-p)/p)]^{1/2}$	$2[(n-p)/n]^{1/2}$	$LD_i(\beta, \sigma^2)=LD_i(\sigma^2)$[20] 5) AP_i statistic is based on confidence ellipsoid and $0\leq AP_i\leq 1$. 6) W_i uses $X'_{(i)}X_{(i)}$ matrix in calculations and more sensitive to the high leverage point then the other statistics. 7) C_i^* statistic is more suitable for the graphical examinations (as normal probability plots) in balanced cases ($h_{ii}=p/n$) and when squared values of C_i^*.

Table 2:(Cont.)

3. Discussion and Classification

An observation may not have the same influence on all regression results and several statistical measures have been proposed in the literature for identifying influential observations in linear regression analysis as it was seen above. A classifications have been done according to base of the measures as a) measures based on residuals, b) measures based on influence curve, c) measures based on volume of confidence ellipsoids, d) measures based on the Likelihood function, e) measures based on the subset of the regression coefficients by [6]. But these classifications are not helpful for the users in practice. Because, there are measures, which based on the same base, but measure the influence on different results of the regression. This situation get mixed up the analyst's mind dial with the subject if it is necessary to examine all of these measures or not. The users usually want to know what the statistics measure, not that the statistics based on. Thus, here it is important that the question "Influence on what?" In this study, the classification approach of the statistics is based on this question's answer.

The primary goal of the analysis may provide the answer to the question of which influence to consider. For example, if $\hat{\beta}$ is of primary concern, then measuring the influence of the observations on $\hat{\beta}$ is appropriate, whereas if prediction is primary goal, then measuring influence on the predicted values may be more appropriate than measuring influence on $\hat{\beta}$. So, it is not necessary to examine all of these statistics for measuring an observation influence. First, It must be selected the statistic among them according to our major concerns about regression results. In that case, the statistics that are used for identifying of influential observations might be classified according to measuring the influence on which results of the regression.

These statistics can be classified according to common properties and differences in Table 2 as follows;

1. The statistics detect outlier.
2. The statistics detect high-leverage points.
3. The statistics measure influence on $\hat{\beta}$.
4. The statistics measure influence on variance (s^2).
5. The statistics measure influence on $\hat{y}$.
6. The statistics measure influence on $\hat{\beta}$ and s^2.

The classification of the statistics according to the approach above, some purpose of the identification of the observation, used statistics and some important results are given in Table 3.

Now, the analyst can select and use the statistics suitable for his/her purpose easier instead of using all statistics. Thus, being much of the statistics will not get mixed up the analyst's mind. In some groups, there were more than one statistics and these statistics might behave different. Because of that, the same observations might be identified as influential by the same group statistics; the different observations might also be influential. That is, an observation might be influential according to one statistics; the same observation might not be influential according to the other statistics. This could be cause a contradiction among the same group statistics. In this situation, it is necessary to compare ability of the same group statistics each other and to identify the best one in the same groups. For this purpose, the values of some observations were changed in data according to definition of the unusual observations and compared ability of the same group statistics each other. In addition to these results, after examination of some important properties of the statistics and literatures, the sensitive statistics are proposed among the statistics placed in same groups (bold in Table 3).

4. Illustrative Example

In this section, we report the results of the statistics explained above for a numerical example. The data which were taken in fisheries faculty from 138 fish, of which ages (Y) and otolith length (X_1) respectively, are used (Figure 2). It was changed the 40^{th}, 80^{th}, and 120^{th} observations in data set according to definition of the unusual observations as A, B and C points in Figure 1 for testing the sensitivity of the statistics and comparing ability of the same group statistics each other. For the data set the regression equation was fitted and the statistics, which were given, before were calculated, the influential observations were identified (Table 4) and concluded.

Observation	The purpose of the identification	Used statistic	Effects on results when omitting from the data
Outlier	-to indicate violation of assumptions-to need transformation on data or not -to identify the sufficiency of the model	e_{in} e_{is} t_i **t_i***	-change intercept -change the variance of regression coefficients -change s^2.
High-Leverage Points	- to identify the observations(with high leverage) far from the center of X space	**h_{ii}** MD_i $WSSD_i$	- change the standard error of the regression coefficients
Influential Observations	-to identify the influence on estimation of $\hat{\beta}$.	LD_i (β) **CVR_i** CW_i C_i	-Change the variance of regression coefficients -Change the regression coefficients.
	-to identify the influence on variance (s^2)	**LD_i (σ^2)**	-change variance (s^2).
	-to identify the influence on $\hat{y}$	**WK_i**	-change $\hat{y}$
	-to identify the influence on estimation of $\hat{\beta}$ and s^2.	AP_i $LD_i(\beta,\sigma^2)$ W_i **C_i***	-change the regression coefficients and variance (s^2).
	-to identify influence of i *th* observation on the j *th* regression coefficient $\hat{\beta}_j$.	**$DFBETAS_{j,i}$**	-change the regression coefficients

Table 3: The classification of the statistics according to measuring the influence on which results of the regression: the observations, the purpose of the identification, used statistics, and some effects on results when omitting from the data in regression analysis. In bold statistics are proposed

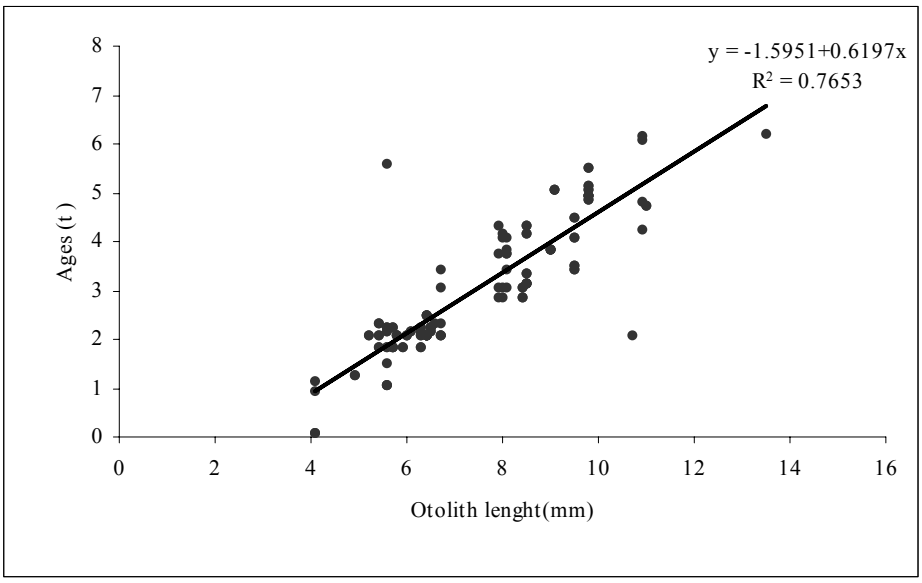


Figure 2: Scatter plot of the data (otolith length versus fish age) and fitted line

Table 4 shows the outliers, high-leverage points and influential observations for the data according to calibration points for all used statistics. Two outliers (observations 40 and 80) and one high-leverage point (observation 120) were detected in 138 data values. The observations 40th 80th and 120th are influential on $\hat{\beta}$ according to statistics measure influence on $\hat{\beta}$. According to the statistics measure influence on variance (s^2), observations 40th and 80th are influential on variance. The observation 105th is influential on $\hat{y}$ in addition to the observations 40th, 80th and 120th, according to the WK_i statistics measure influence on $\hat{y}$. According to the statistics measure influence on $\hat{\beta}$ and s^2, the observations 63th, 64th, and 105th are influential on $\hat{\beta}$ and s^2 in addition to the observations 40th, 80th and 120th. The results of the $DFBETAS_{j.i}$ statistics shows that the observations 3th, 4th, 40th, 61th, 63th, 64 th, 80th, 105th and 120th are influential on $\hat{\beta}_j$.

Observation	Used Statistic	Calibration Point	Influential Observation Numbers
Outlier	e_{in}	1.9021	40,80
	e_{is}	1.9021	40,80
	t_i	1.9700	40,80
	t_i*	**1.9700**	**40,80**
High-Leverage Points	**h_{ii}**	**0.0360**	**120**
	MD_i	0.0360	120
	$WSSD_i$	0.0360	120
Influential Observations	LD_i (β)	0.1026	40,80
	CVR_i	**0.0435**	**40,80,120**
	CW_i	---	40,80
	C_i	0.0513	40,80
	LD $_i$ (σ^2)	**0.1026**	**40,80**
	WK_i	**0.2400**	**40,80,105,120**
	AP_i	0.9565	40,63,80,120
	$LD_i(\beta,\sigma^2)$	0.3518	40,80
	W_i	4.2426	40,80
	C_i*	**1.9854**	**40,63,64,80,105,120**
	$DFBETAS_{0,i}$	**0.1703**	**3,4,40,63,64,80,105,120**
	$DFBETAS_{1,i}$	**0.1703**	**3,4,40,61,63,64,80,105,120**

Table 4: Influential observation numbers according to the different statistics in data. In bold statistics are proposed.

It was seen from the result in Table 4 that the observations that we changed their values for testing the sensitivity of the statistics in the data are generally determined by all of the statistics. Although LD_i (β), CW_i and C_i are influenced from the outliers more than the statistic CVR_i , they are not influenced from high-leverage points among the statistics measure influence on $\hat{\beta}$. The statistic CVR_i is influenced from both outliers and high-leverage points differently. Both LD(σ^2) and WK_i statistics are influenced from both outliers and high-leverage points similarly. From the statistics group measure influence on $\hat{\beta}$ and s^2, although the $LD_i(\beta,\sigma^2)$ and W_i, are influenced from the outliers more than the statistics AP_i and C_i* similarly, they are not influenced from high-leverage points. The statistics AP_i and C_i* are influenced from both outliers and high-leverage points differently.

The statistic CVR_i from the statistics measure influence on $\hat{\beta}$ and the statistic C_i^* from the statistics measure influence on $\hat{\beta}$ and s^2 are more sensitive to unusual observations than the others because of influenced from both outliers and high-leverage points.

If your purpose is	to identify outlier	Then it might used	t_i^*
	to identify high-leverage point		h_{ii}
	to identify the influence of an observation on regression coefficients		CVR_i
	to identify the influence of an observation on variance(σ^2)		$LD_i(\sigma^2)$
	to identify the influence of an observation on predicted values		WK_i
	to identify the influence of an observation on both regression coefficients(β) and variance(σ^2)		C_i^*
	to identify the influence of i *th* observation on j *th* β		$DFBETAS_{j,i}$

Table 5: The proposed statistics according to the purpose.

5. Conclusion

There are three main stages for fitting a regression model by least squares regression analysis. i) The identification of the data quality for a proposed model, ii) the model quality for a given data set, and iii) a fulfillment of all least squares assumptions. Examination of the influential observations has an important place as a diagnostic strategy in all of the stages. Here, it might be suggested to use given in Table 5 for identifying of influential observations according to your purpose among the statistics given above;

Many authors also suggested the same statistics for these purposes [5], [6], [19].

As a result; it should be noted that, outliers or high leverage points should not be automatically rejected but rather should receive special attention and careful examination to determine the cause of their peculiarities. If these points are truly genuine observations, they may indicate violation of assumptions and perhaps the need for an alternative model.

REFERENCES

[1] SHAPIRO, S.S., WILK, M.B. (1965) *An Analysis of Variance Test for Normality. Biometrika*, 52, 591-611.

[2] WONNACOTT, T. H., WONNACOTT, R. J. (1981) *Regression: A second Course in Statistics.* Wiley. New York.

[3] DRAPER, N. R., SMITH, H. (1981) *Applied Regression Analysis. John Willey and Sons,* Inc. New York.

[4] BELSLEY, D.A., KUH,E., WELSCH,R.E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Colinearity.* Wiley. New York.

[5] CATTERJEE, S., HADI, A.S. (1986) *Influential Observation, High Leverage Points, and Outliers in Linear Regression.* Statistical Science, 1, 379-393.

[6] CATTERJEE, S., HADI, A.S. (1988) *Sensitivity Analysis in Linear Regression.* John Willey and Sons Inc. Canada.

[7] MELOUN, M., MILITKY, J. (2001) *Detection of Influential Points in OLS Regression Model Building.* Analytica Chimica Acta, 439, 169-191.

[8] ATKINSON, A.C. (1981) *Two Graphical Displays for Outlying and Influential Observations in Regression.* Biometrika, 13-20.

[9] MYERS, R.H. (1990) *Classical and Modern Regression with Applications.*Duxbury. New York.

[10] HOAGLIN, D.C., WELSCH, R.E. (1978) *The Hat Matrix in Regression and ANOVA.* The American Statistician, 32, 17-22.

[11] VELLEMANN, P.F., WELSCH, R.E. (1981) *Efficient Computing of Regression Diagnostics.* The American Statistician, 35, 234-242.

[12] FARBER, O., KADMON, R. (2003) *Assessment of Alternative Approaches for Bioclimatic Modeling with Special Emphasis on the Mahalanobis Distance.* Ecological Modelling, 160, 115-130

[13] FERNANDEZ Pierna, J. A., WAHL, F., de Noord, O. E., MASSART, D. L. (2002) *Methods for Outlier Detection in Prediction.* Chemometrics and Intelligent Laboratory Systems, 63, 27-39.

[14] COOK R.D. (1977) *Detecting of Influential Observations in Linear Regression Technometrics,* 19, 15-18.

[15] McDONALD, B. (2002) *A teaching Note on Cook's Distance-A Guideline.* Res.Lett. Inf. Math. Sci., 3, 127-128.

[16] MONTGOMERY D.C., PECK, E.A. (1992) *Introduction to Linear Regression Analysis.* John Willey and Sons, Inc. Canada.

[17] COOK, R.D., WEISBERG, S. (1982) *Residuals and Influence in Regression.* Chapmann and Hall. London.

[18] WELSCH, R.E. (1982) *Influence Functions and Regression Diagnostics in Modern Data Analysis* (R.L. Launer and A.F. Siegel, eds.). Academic Press. New York.

[19] HOAGLIN, D.C. KEMPTHORNE, P.J. (1986) *Comment. Statistical Science, 1,* (3), 408-412.

[20] COOK, R.D. (1986) *Comment. Statistical Science, 1,* No.3: 393-397.