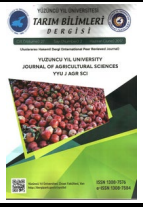
Yüzüncü Yıl Üniversitesi
Tarım Bilimleri Dergisi
(YYU Journal of Agricultural Science)

http://dergipark.gov.tr/yyutbd



Araştırma Makalesi (Research Article)

Detection and Diagnostic Methods of Multiple Influential Points in Binary Logistic Regression Model in Animal Breeding

Burcu MESTAV[*1]

[1]Çanakkale Onsekiz Mart Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, 17020, Çanakkale , Türkiye
*Sorumlu yazar e-posta:burcumestav@comu.edu.tr

Article Info

Geliş: 25.10.2019
Kabul: 28.11.2019
Online Yayınlanma 31.12.2019
DOI: 10.29133/yyutbd.638226

Keywords

Multiple Influential Point, Generalized Standardized Pearson Residual (GSPR), Generalized Difference of Fits (GDFFITS) Generalized Square Difference in Beta (GSDFBETAS)

Abstract: Multiple influential points adversely affect parameter estimation in binary logistic regression models and lead to misinterpretation of results. An influential point is a data point that does not follow the overall slope of remaining data and has extreme value in terms of x. Since the presence of approximately 10% of influential points in a dataset affects parameter estimates, detection and diagnosis of these points greatly matter. Graphical (such as scatter graph and box graph) and analytical methods are adopted in the detection and diagnosis of multiple influential points. Among the commonly used diagnostic methods are Pearson residuals, Standardized Pearson Residuals (SPR), Cook Distance (CD), Hat matrix, DFFITS, and DFBETA. However, these methods mask problems and fail to diagnose if there are multiple influential points. Many statisticians have developed and proposed new diagnostic methods, such as Generalized Standardized Pearson Residual (GSPR) and Generalized Weights (GW), to overcome this problem. This study exploited a dataset containing multiple influential points (15%) for weaning weight (WW), yearling weight (YW), fleece weight (FW), and fertility rate (FR) of Romney ewes and modelled the effects of WW, TW and FW variables on FR by binary logistic regression model. This study is intended to determine the multiple influential points by graphical methods and to examine the performance of commonly used and newly developed methods in the diagnosis of these data points. As a result, it was observed that the commonly used methods mask multiple influential points and the new proposed methods competently identify these points.

Hayvan Islahında İkili Lojistik Regresyon Modelinde Çoklu Etki Noktalarının Tespit ve Teşhis Yöntemleri

Makale Bilgileri

Received: 25.10.2019
Accepted: 28.11.2019
Online Published 31.12.2019
DOI: 10.29133/yyutbd.638226

Anahtar kelimeler
Çoklu Etkili Gözlem Noktası, Genelleştirilmiş Standartlandırılmış Pearson Artığı (GSPA), Genelleştirilmiş Uyum Farkı (GDFFITS)

Öz: Çoklu etkili gözlem noktaları ikili lojistik regresyon modellerinde parametre tahminlerini olumsuz yönde etkilemekte ve sonuçların yanlış yorumlanmasına sebep olmaktadır. Bir etkili gözlem noktası verilerin geri kalanının genel eğimini takip etmeyen ve x bakımından aşırı değere sahip olan bir veri noktasıdır. Veri seti içinde yaklaşık % 10 ve üzerinde etkili gözlem noktasının bulunması parametre tahminlerini etkilediği için bu noktaların tespit ve teşhisi oldukça önemlidir. Çoklu etkili gözlem noktalarının tespit ve teşhisinde grafiksel (saçılım grafiği ve kutu grafiği gibi) ve analitik yöntemler kullanılmaktadır. En yaygın kullanılan teşhis yöntemleri Pearson Artıklar, Student Türü Artıklar, Şapka Matrisi, Cook Uzaklığı, DFFITS, DFBETA vb. yöntemlerdir. Ancak bu yöntemler çoklu etkili gözlem noktalarının olması durumunda maskeleme problemleri ile karşılaşmakta ve teşhiste başarısız olmaktadır. Bir çok istatistikçi bu problemle başedebilmek için Genelleştirilmiş Standartlandırılmış Pearson

Betadaki Genelleştirilmiş Kare Farkı (GSDFBETA)

Artığı (GSPA), Genelleştirilmiş Ağırlıklar (GA) gibi yeni yöntemler geliştirmiş ve önermiştir. Bu çalışmada, Romney ırkı koyunlardan elde edilen sütten kesim ağırlığı (SKA), Bir yaş canlı ağırlığı (BYCA), yapağı ağırlığı (YA) ve doğurganlık oranı (DO) değişkenlerine ait içinde çoklu etkili gözlem noktası (%15) bulunan veri seti ile çalışılmış ve DO üzerine SKA, BYCA ve YA değişkenlerinin etkisi ikili lojistik regresyon modeli ile modellenmiştir. Çalışmanın amacı çoklu etkili gözlem noktalarını grafiksel yöntemlerle tespit edip yaygın olarak kullanılan ve yeni geliştirilmiş yöntemlerin bu veri noktalarının teşhisindeki performanlarını incelemektir. Çalışmanın sonucunda yaygın olarak kullanılan yöntemlerin çoklu etkili gözlem noktalarını maskelediği ancak yeni önerilen yöntemlerin bu noktaları başarılı şekilde teşhis ettiği gözlenmiştir.

1. Introduction

The Binary Logistic Regression (BLR) model has been commonly used in the analysis of the functional relationship between an outcome variable and predictor variable(s) in animal breeding for many years and studied by a great number of researchers in recent years (Eyduran et al., 2005; Gaskins et al., 2005; Korkmaz et al., 2012; Aktaş and Doğan, 2014; Yakubu et al., 2014; Aktaş et al., 2015; Takma et al., 2016; Erdinç et al., 2017; Baeza-Rodriguez et al., 2018; Gebre et al., 2018). The most important difference of BLR from the general linear regression model is that the outcome variable refers to binary outcome which is assigned 0 or 1. Therefore, the error variance becomes nonconstant and the error term exhibits logistic distribution. BLR assumes that the sample size is adequate and a high correlation among the predictor variables does not exist and lastly there should be no outlier and/or influential point in the dataset (Hilbe, 2009). An unknown parameter in BLR is estimated by using maximum likelihood (ML), but it is well known that ML can be severely affected in the presence of outliers. The outliers are named differently according to their position on the X and Y axis. For example, both outliers and influential points are measurements that do not fit in the trend shown by the rest of the data. Hence, these two concepts should not be mistaken for each other. To specify, an outlier is an unusual observation whose outcome y does not follow the general slope of the rest of the data, whereas the influential point is a data point that does not follow the general slope of the rest of the data and has an extreme predictor x value. Parameter estimates obtained in the presence of influential points, in particular, will cause misinterpretation of the results. Moreover, the binary outcomes are likely to be misclassified. Hampel et al. (1986) have claimed that if these outliers occur in about 1-10% of the dataset, it is normal and can be removed from the dataset; however, if there are more than 10% outliers, it is recommended to use a robust estimator instead of ML estimator (Midi and Ariffin, 2013). Outliers and influential points often cause problems in the analyses of data in animal and plant breeding. Some researchers have reported that performance of accuracy estimation in genomic prediction methods used in genomic selection studies is adversely affected by outliers (Via et al. 2012; Heslot et al., 2013; Estaghvirou et al., 2014). Therefore, the detection of outliers or influential points is crucial and must be performed before the analysis. Result of a diagnosis refers to a specific amount that is computed from the data and calculated to determine the influential points where the influential points can be eliminated or corrected. Thus, such observations need to be described and their effects on the model and subsequent analysis should be investigated (Nurunnabi et al., 2010). In recent years, diagnostic and detection have become an almost indispensable part of BLR and a great many statisticians have studied diagnostic and detection methods of outliers and/or influential points. Before Imon and Hadi (2008), the diagnostic methods always relied on the detection of outlier. However, the subsequent studies have shown that the observation points that cause significant deviation in parameter estimates are influential points. Since the influential point too is an outlier, the diagnostic methods before Imon and Hadi (2008) are valid for influential points. However, the general objective of all the new diagnostic methods including Imon and Hadi (2008) and the subsequent ones is to detect multiple influential points. Diagnosis of outliers and/or influential points based on residuals is known as the most common method in BLR (Pregibon, 1981; Jennings, 1986; Copas, 1988). The most commonly employed diagnostic methods for the identification of outliers in BLR are Pearson residuals, Standardized Pearson Residuals (SPR), Cook Distance (CD), Hat matrix, Difference of Fits (DFFITS), Difference in Beta (DFBETA). However, these methods are only able to

identify single outliers. If the dataset contains multiple outliers/influential points, these methods fail to identify them because of the masking and swamping problems (Imon and Hadi, 2008; Habshah et al., 2009; Sanizah et al., 2011). Recently, diagnostic methods have been developed by a great number of statisticians to overcome these problems (Cook, 1977; Pregibon; 1981, Jennings; 1986, Copas, 1988; Hadi and Simonoff, 1993; Imon, 2006; Imon and Hadi, 2008; Habshah et al., 2009; Nurunnabi et al., 2010; Sarkar et al., 2011). The new approaches developed based on a deleted group are Generalized Standardized Pearson Residual (GSPR), Generalized Weights (GW), Generalized Difference of Fits (GDFFITS), and Generalized Square Difference in Beta (GSDFBETAS). Studies have shown that these methods successfully cope with masking and swamping problems in datasets with multiple influential points. The prediction of genetic parameters and accuracy of breeding values are greatly important for animal breeding and animal improvement programs. In addition, parameter estimation of risk factors affecting some economically important traits, such as fertility rate, birth type, the stillbirth rate, in terms of care and management plays a critical role in the livestock field. Influential points adversely affect the achievement of parameter estimates of traits. However, to date, their detection in animal breeding has not yet been evaluated. Therefore, it has become necessary to identify influential points in the datasets in order to obtain accurate parameter estimates. Accordingly, the aim of this study is to contribute to these scholarly efforts by introducing various existing diagnostic and detection methods adopted to identify multiple influential points in a dataset analyzed by using BLR in animal breeding.

2. Materials and Methods

2.1. Material

Animal materials of this study consisted of 100 Romney ewes raised in New Zealand. Since the aim of the study was to compare multiple influential points and diagnostic methods, the dataset was arranged to contain 15% influential points. Of the 100 units of data, 85 were selected from the 300 units of data using the random sampling method, while 15 were the influential points already present in the dataset. Thus, the dataset with 15% influential point was created. The study was conducted over this dataset.

2.2. Method

Binary Logistic Regression model was used to determine the influence of weaning weight (WW), yearling weight (YW) and fleece weight (FW) of the ewes on fertility rate (FR). Binary variable was coded as 1 (lambed) or 0 (unlambed) in relation to FR. The mathematical model of BLR was as follows:

$$Y = \pi(x) + \varepsilon \tag{1}$$

where Y is an nx1 vector of the outcome variable (FR), which is denoted by $y = 1$ or 0 with probabilities π and $1-\pi$, respectively. ε is a nx1 vector of error terms:

$$\varepsilon = \begin{cases} 1-\pi & \text{with probability } \pi \quad \text{if } y = 1 \\ -\pi & \text{with probability 1-}\pi \quad \text{if } y = 0 \end{cases} \tag{2}$$

which follows a distribution with mean zero and variance $\pi(1-\pi)$.

$$\pi(x) = E(Y \mid X) = \exp(x^T\beta) \Big/ \left[1 + \exp(x^T\beta)\right], 0 \le \pi(x) \le 1 \tag{3}$$

with $\beta^T = (\beta_0, \beta_1, \beta_2, \beta_3)$ being the vector parameters, X is an nxk ($k = p+1$) matrix of predictor variables (WW, YW, and FW) and is non-linear in $\pi(x)$. Thus, we have to use the logit link function to transform it into a linear form.

In the literature, there are many methods of detection and diagnostic of influential points. All of these methods were developed firstly for general linear regression and then they were suggested for BLR by Pregibon (1981). They can be divided into two groups: e.g., graphical and analytical methods. The best-known graphical methods are the scatter, box, and residual plots. However, since the graphical methods fail to provide reliable information, the analytical methods are preferred, especially when the number of predictor variables is high. Many analytical methods are proposed in the related literature. In this study, the most commonly used analytical methods were adopted which were thought to prove more useful for researchers in animal breeding. The analytical methods are statistical values computed from the dataset that can be used to identify the presence of influential points. Although the main tools of the developed analytical methods are residuals, the methods having been developed in recent years are based on the deletion of suspected observations. In BLR, the primary building blocks of analytical methods used to identify influential points are residual vector and projection (leverage) matrix (Pregibon, 1981). According to a similar approach to linear regression (Copas, 1988), the i^{th} residual is defined in BLR as follows:

$$r_i = y_i - \hat{\pi}_i, \quad i = 1,...,n \tag{4}$$

Although residual, also known as raw residual, is very important in detecting ill-fitting, residuals defined in equation (3) are unscaled. Therefore, it is not applicable to influential points diagnosis. There are two versions of the scaled residual type commonly used in BLR to eliminate this problem: Pearson Residuals (PR) and Standardized Pearson Residuals (SPR). PR can be defined as:

$$r_{Pi} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i\left(1-\hat{\pi}_i\right)}}, i = 1,...,n \tag{5}$$

Pearson residual value of an observation is considered a residual outlier if it's greater than 3 by absolute value (Ahmad et al., 2011). Standardized Pearson Residuals value is obtained by dividing the raw residuals by the standard error provided by $se\left(y_i - \hat{\pi}_i\right) = \sqrt{v_i\left(1-h_{ii}\right)}$, where $v_i = \hat{\pi}_i\left(1-\hat{\pi}_i\right)$ and h_{ii} is the i^{th} diagonal element of the nxn matrix, known as hat matrix, $H = V^{1/2}X\left(X^TVX\right)^{-1}X^TV^{1/2}$. If $h_{ii} > \{2 \text{ or } 3\}k/n$, then this may evidence the presence of influential points (Friendly and Meyer, 2015). V is a diagonal matrix with diagonal elements v_i (Pregibon, 1981). Hence, the SPR for BLR can be defined as:

$$r_{Si} = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i\left(1-h_{ii}\right)}}, i = 1,...,n \tag{6}$$

In BLR, observations with SPRs, which are less than -3 and greater than +3, are considered as outliers (Midi and Ariffin, 2013). Methods other than the methods of identifying influential points/outliers using residuals delete suspect observations. The most common diagnostic statistics adopting observation deletion approach are Cook Distance (CD), Hat matrix (Lev), Difference of Fits (DFFITS) and Difference in Beta (DFBETA) (Cook, 1977; Belsley et al.,1980; Nurunnabi et al., 2010). Pregibon (1981) defined CD by using linear regression models for BLR as follows:

$$CD_i = \frac{r_{Si}^2 h_{ii}}{k\left(1-h_{ii}\right)}, i = 1,...,n \tag{7}$$

If there is an observation with the value of $CD_i > 1$, it is regarded as an influential point. Another influential point determination measure similar to CD is DFFITS value suggested by Welsch (1982). DFFITS is defined in terms of SPR and Lev values in BLR as follows:

$$DFFITS_i = r_{Si}\sqrt{\frac{h_{ii}}{(1-h_{ii})}\frac{v_i}{v_{(-ii)}}}, i = 1,...,n \tag{8}$$

An influential point has DFFITS > 2 or $3\sqrt{k/n-k}$.

Although the abovementioned methods often provide effective results in determining influential points, they are effective if there is only one single influential point in the dataset. If there are multiple influential points in the dataset, these methods are ineffective. In the case of multiple influential points, they cause masking and swamping problems (Imon and Hadi, 2008). Therefore, new approaches are needed to prevent these problems from occurring. The proposed approaches based on a deleted group in the BLR are Generalized Standardized Pearson Residual (GSPR), Generalized Difference of Fits (GDFFITS), and Generalized Square Difference in Beta (GSDFBETAS) (Imon and Hadi, 2008; Nurunnabi et al., 2010; Nurunnabi and Nasser, 2011). These methods have been obtained by generalizing the existing methods and are based on deletion of the suspected group from the dataset (Hadi and Simonoff, 1993). Before using these methods, the dataset is examined using scatter plot and possible influential points are identified. Then, the d-dimensional observations, which are considered influential points in the n-dimensional dataset, are deleted before the fitting of the model. R and D, respectively, represent the set of situations of the "remaining" and "deleted" observations. The parameters of the model with the remaining set are estimated by using ML. Statistical values of the proposed methods were obtained with estimated parameters. Thus, the probability values determined according to the R set are defined as:

$$\hat{\boldsymbol{\pi}}_i^{(-D)} = \frac{exp\left(x_i^{'}\hat{\boldsymbol{\beta}}^{(-D)}\right)}{1+exp\left(x_i^{'}\hat{\boldsymbol{\beta}}^{(-D)}\right)}, i = 1,...,n \tag{9}$$

In this case, after the ith observation is deletion, the residuals are defined as follows:

$$\hat{\boldsymbol{\varepsilon}}_i^{(-D)} = y_i - \hat{\boldsymbol{\pi}}_i^{(-D)}, i = 1,...,n \tag{10}$$

The variance and leverage values of the observation set in question are computed by the following equations:

$$v_i^{(-D)} = \hat{\boldsymbol{\pi}}_i^{(-D)}\left(1-\hat{\boldsymbol{\pi}}_i^{(-D)}\right) \tag{11}$$

$$h_{ii}^{(-D)} = \hat{\boldsymbol{\pi}}_i^{(-D)}\left(1-\hat{\boldsymbol{\pi}}_i^{(-D)}\right)x_i^T\left(X_R^T V_R X_R\right)^{-1}x_i \tag{12}$$

The proposed GSPR value is obtained by following equations using equations 9, 10, 11, and 12 (Nurunnabi and West, 2012).

$$r_{si}^{(-D)} = \begin{cases} \dfrac{y_i - \hat{\boldsymbol{\pi}}_i^{(-D)}}{\sqrt{v_i^{(-D)}\left(1-h_{ii}^{(-D)}\right)}}, \text{for } i \in \text{R} \\ \dfrac{y_i - \hat{\boldsymbol{\pi}}_i^{(-D)}}{\sqrt{v_i^{(-D)}\left(1+h_{ii}^{(-D)}\right)}}, \text{for } i \in \text{D} \end{cases} \tag{13}$$

An observation is described as an influential point when its corresponding GSPR value of any observation is 3 points greater than the absolute value. The GDFFITS method suggested by Nurunnabi et al. (2010) is defined in (14) using (13):

$$GDFFITS = r_{si}^{(-D)}\sqrt{h_{ii}^{*(-D)}} \tag{14}$$

where, $h_{ii}^{*(-D)} = \begin{cases} \dfrac{h_{ii(R)}}{1-h_{ii(R)}} \text{ for } i \in \mathrm{R} \\ \dfrac{h_{ii(R)}}{1+h_{ii(R)}}, \text{for } i \in \mathrm{D} \end{cases}$.

If the GDFFITS value corresponding to the ith observation is $\left|GDFFITS_i\right| \geq 2$ or $3\sqrt{\dfrac{p+1}{n-d}}$, it means that the observation is the influential point (Nurunnabi et al., 2010). Another proposed method for diagnostic of the influential point is GSDFBETA method, suggested by Nurunnabi and Nasser (2011). The GSDFBETA is defined as:

$$GSDFBETA = \begin{cases} \dfrac{h_{ii}^{*(-D)} r_{si(R)}^2}{1-h_{ii(R)}}, \text{for } i \in \mathrm{R} \\ \dfrac{h_{ii}^{*(-D)} r_{si(R)}^2}{1+h_{ii(R)}}, \text{for } i \in \mathrm{D} \end{cases} \tag{15}$$

If $\left|GSDFBETA_i\right| \geq \dfrac{\left(3\sqrt{p+1/n-d}\right)^2}{1-\left[3p/n-d\right]}$, i^{th} observation is considered as influential point.

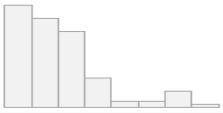
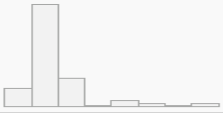
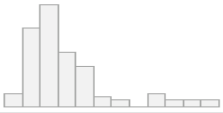
To detect influential points, the dataset was analyzed with Maximum Likelihood (ML) and then the diagnostic and detection methods were analyzed. R version 3.5.1 (R Development Core Team 2018) software was used for both analysis and detection.

3. Results

Descriptive statistics and histogram graphs of the predictor variables used in the study are given in Figure 1. Histogram graphs in Figure 1 show that the distributions of the predictor variables are skewed, and outliers have an effect on the dataset.

Data Frame Summary

Dimensions: 100 x 3
Duplicates: 3

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	WW [numeric]	Mean (sd) : 25.5 (8.4) min < med < max: 15 < 25 < 51.3 IQR (CV) : 10 (0.3)	40 distinct values		100 (100%)	0 (0%)
2	YW [numeric]	Mean (sd) : 3.9 (1.3) min < med < max: 2.5 < 3.7 < 9.8 IQR (CV) : 0.8 (0.3)	44 distinct values		100 (100%)	0 (0%)
3	FW [numeric]	Mean (sd) : 27.5 (12.2) min < med < max: 13 < 23 < 69 IQR (CV) : 11.2 (0.4)	47 distinct values		100 (100%)	0 (0%)

Generated by summarytools 0.9.4 (R version 3.5.1)

Figure 1. Descriptive statistics and histogram graphs of the predictor variables

The most common graphical methods used to determine whether there are outliers in the dataset before analysis are scatter plot and box plot. The scatter plots of FR against WW, YW and FW is shown in Figures 2 and 3. The plots evidence the presence of suspicious observations (between Observations 85 and 100) that can be regarded as multiple influential points.

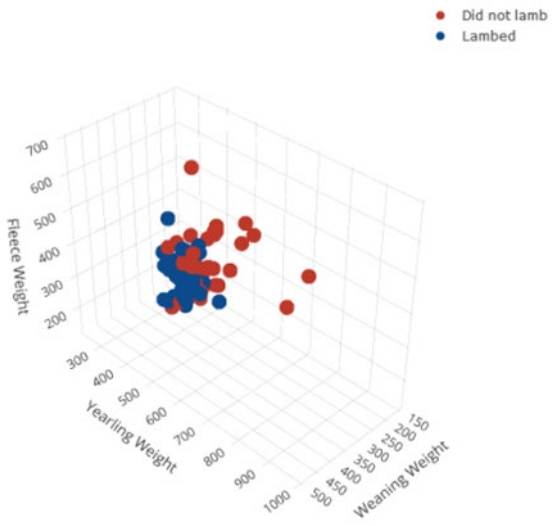


Figure 2. Scatter plot of Fertility Rate against Weaning Weight, Yearling Weight and Fleece Weight.

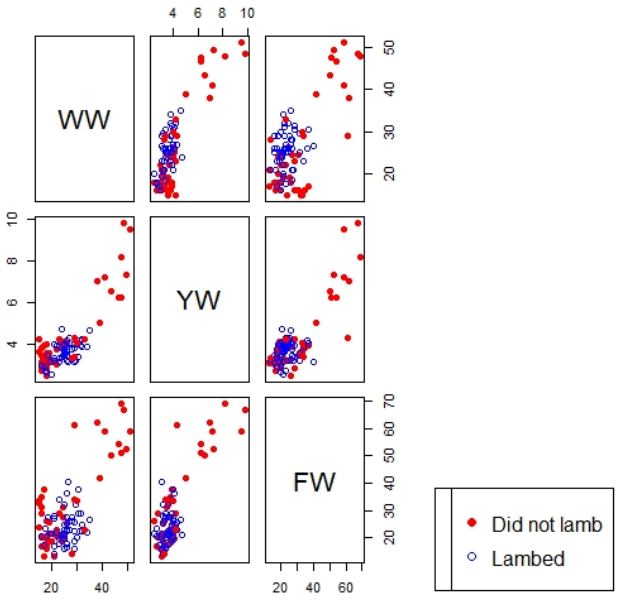


Figure 3. Scatter plots of FR on two predictors.

The box plots of each predictor are given in Figure 4. It shows that many observation points can be outliers, as in Figures 2 and 3.

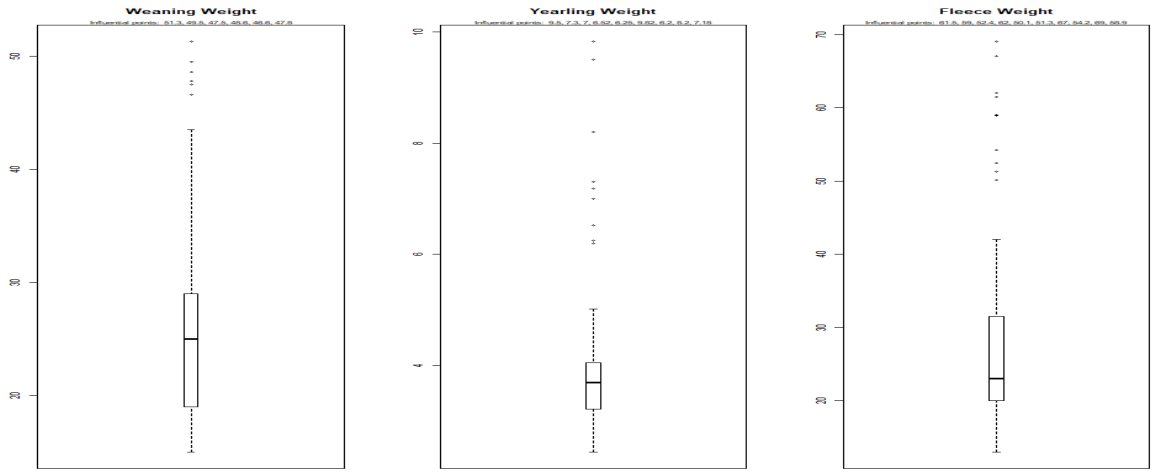


Figure 4. Box plots of WW, YW and FW

The plots of FR against WW, YW, and FW clearly present that the observations between 85 and 100 may severely distort the covariate pattern. However, scatter plots and box plots alone are incompetent at the diagnosis of suspicious observations. For this reason, we need analytical methods to determine the extent of the influence of suspicious observations determined by graphical methods. Until the development of new approaches, diagnostic methods (CD, PR, SPR and DFFITS) worked functionally in the presence of a single outlier, whereas they were inadequate when multiple influential points were observed. Table 1 shows the CD, PR, SPR, and DFFITS results of suspicious observations graphically detected in Figures 2, 3, and 4. Table 1 reveals that the degree of other suspected observations' effects on the dataset except for the 99th observation value in PR and SPR is below the cut-off limit and the most commonly used one of these diagnostics fail to determine the influential points. It seems that CD (using 1 as the cut-off value) and DFFITS (using 0.613 as the cut-off value) fail to determine any influential points in the dataset, whereas PR (using 3 as the cut-off value) and SPR (using 3 as the cut-off value) can correctly determine only the 99th observation as influential point. The index plot in Figure 5 shows that PR and SPR can correctly and clearly determine the influential point compared to CD and DFFITS. This is due to the masking problem of these methods when there are multiple influential points. The use of these methods may mislead researchers and continuing the analysis without removing suspicious observation points may lead to misinterpretation of parameter estimates.

Table 1. Results of Cook Distance, Pearson Residuals (PR), Standardized Pearson Residuals (SPR) and Difference of Fits (DFFITS),

Number of Observation	CD (1.00)	PR (3.00)	SPR (3.00)	DFFITS (0.613)	Number of Observation	CD (1.00)	PR (3.00)	SPR (3.00)	DFFITS (0.613)
1	0.000	0.224	0.228	-0.057	.	.	.	.	.
2	0.005	0.939	0.950	-0.164	86	0.004	0.318	0.339	-0.166
3	0.003	0.471	0.485	-0.147	87	0.000	0.069	0.069	-0.013
4	0.001	0.277	0.282	-0.072	88	0.012	0.445	0.487	-0.276
5	0.001	0.267	0.272	-0.070	89	0.000	0.098	0.099	-0.017
6	0.005	1.006	1.016	-0.159	90	0.007	0.442	0.468	-0.207
7	0.002	0.375	0.383	-0.104	91	0.052	0.876	0.968	-0.525
8	0.005	0.741	0.754	-0.173	92	0.000	0.028	0.028	-0.002
9	0.015	1.118	1.143	-0.262	93	0.034	1.190	1.241	-0.391
10	0.013	1.154	1.175	-0.243	94	0.035	0.738	0.813	-0.447
11	0.006	0.709	0.725	-0.189	95	0.000	0.094	0.095	-0.019
12	0.011	1.190	1.208	-0.222	96	0.000	0.136	0.138	-0.032
13	0.003	0.471	0.485	-0.147	97	0.062	1.857	1.919	-0.442
14	0.001	0.277	0.282	-0.072	98	0.048	2.191	2.233	-0.358
15	0.001	0.267	0.272	-0.070	**99**	0.090	**3.446**	**3.496**	-0.379
.	.	.	.	.	100	0.012	1.185	1.203	-0.229

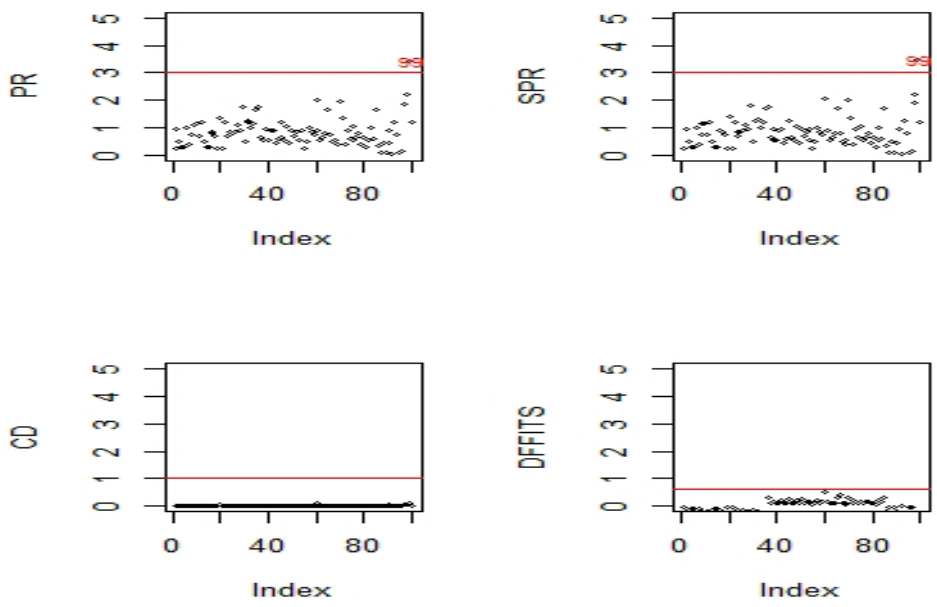


Figure 5. Index plots of CD, PR, SPR, and DFFITS for dataset

The results of GSPR, GSDFFITS, and GSDFBETA methods proposed as new approaches in this study are available in Table 2. It is clear from the table that the GSPR, GSDFFITS, and GSDFBETA values for the suspected observations were much larger than the others and all exceeded the cut-off values of 3.00, 0.651, and 0.474, respectively. The advantage of these methods over other methods is that they are robust to the masking problem. Similar conclusions may be drawn from the index plots of GSPR, GSDFFITS, and GSDFBETA as presented in Figure 6. All these 15 suspected observations are separated from the other data and correctly determined as influential points.

Table 2. Results of GSPR, GSDFFITS, and GSDFBETA

Number of Observation	GSPR (3.00)	GSDFFITS (0.651)	GSDFBETA (0.474)	Number of Observation	GSPR (3.00)	GSDFFITS (0.651)	GSDFBETA (0.474)
1	0.114	0.016	0.000	.	.	.	.
2	0.738	0.131	0.018	**86**	**4.438**	**1.816**	**2.746**
3	0.286	0.063	0.004	**87**	**8.291**	**3.149**	**8.484**
4	0.221	0.045	0.002	**88**	**40.321**	**2.745**	**7.502**
5	0.225	0.048	0.002	**89**	**3.005**	**1.702**	**1.967**
6	2.087	0.417	0.181	**90**	**18.583**	**2.255**	**5.008**
7	0.320	0.074	0.006	**91**	**64.356**	**2.518**	**6.332**
8	0.384	0.074	0.006	**92**	**2.708**	**1.906**	**1.834**
9	0.481	0.134	0.019	**93**	**22.520**	**1.721**	**2.945**
10	0.543	0.143	0.022	**94**	**56.042**	**2.558**	**6.528**
11	0.403	0.090	0.008	**95**	**13.872**	**3.222**	**9.823**
12	0.661	0.166	0.029	**96**	**5.690**	**2.210**	**4.146**
13	0.286	0.063	0.004	**97**	**8.291**	**1.150**	**1.297**
14	0.221	0.045	0.002	**98**	**8.828**	**0.960**	**0.911**
15	0.225	0.048	0.002	**99**	**5.335**	**0.854**	**0.710**
.	.	.	.	**100**	**5.500**	**0.904**	**0.795**

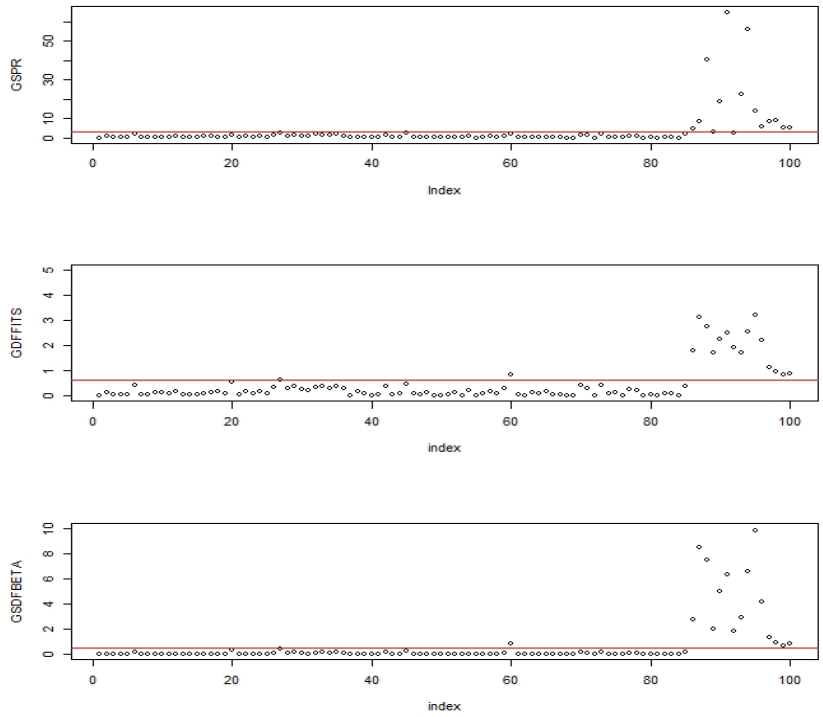


Figure 6. Index plots of GSPR, GSDFFITS, and GSDFBETA for the dataset.

It is crucial for researchers when analyzing data to be able to determine influential points. The results of the analysis with ML estimator of the dataset with influential points and without influential points are presented in Table 3. In Table 3 on the results of both datasets, it was observed that FW variable had no statistically significant contribution to FR ($p>0.10$), whereas WW and YW variables contribute significantly to FR ($p<0.05$). Furthermore, the coefficients of the dataset without influential points and the dataset with influential points differ. As a result, researchers can remove observations that they detect, both graphically and using suggested methods, from the dataset by looking at the size of the dataset and the percentage of influential points in the dataset.

Table 3. Results of the analysis of the dataset with influential points and without influential points.

Analysis of the dataset with influential points					
Coefficients	Estimate	Std. Error	z value	p-value	
(Intercept)	2.053	1.124	1.826	0.068	.
WW	0.025	0.007	3.831	0.000	***
YW	-0.017	0.006	-2.964	0.003	**
FW	-0.007	0.004	-1.879	0.060	.
Analysis of the dataset without influential points					
Coefficients	Estimate	Std. Error	z value	p-value	
(Intercept)	-4.512	2.266	-1.991	0.047	*
WW	0.049	0.011	4.519	0.000	***
YW	-0.018	0.008	-2.263	0.024	**
FW	0.002	0.006	0.412	0.681	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

4. Discussion and Conclusion

The aim of this study is to comparatively examine the performances of the detection and diagnostic methods (graphical and analytical methods) used in the presence of multiple influential points in a dataset where the effect of WW, YW, and FW variables on FR is modeled. Outlier/influential points occur in almost all research studies and this type of observations are a problem in statistical analysis. Therefore, their detection and diagnosis are a crucial issue that needs to

be addressed before further analysis is performed. Analysis without determining the location and amount of these observation points adversely affects parameter estimates, particularly data analysis with outliers/influential points. Results of a breeding program and management strategy plan to be carried out using the predicted parameters may differ from the expected outcome, which in turn directly affects the economic situation. Therefore, it is necessary to identify influential points in the datasets in order to obtain accurate parameter estimates. In this study, four of the most commonly used methods and three novel methods for the diagnosis of multiple influential points in BLR are introduced and their performances are comparatively examined. Evaluating the diagnostic methods in terms of performance shows that the proposed method (GSPR, GSDFFITS, and GSDFBETA) is highly competent at determining multiple influential points in the case of failure of the existing commonly used diagnostic methods (CD, PR, SPR, and DFFITS).

References

Aktaş, A. H. & Doğan, Ş. (2014). Effect of live weight and age of Akkaraman ewes at mating on multiple birth rate, growth traits, and survival rate of lambs. *Turk. J. Vet. Anim. Sci., 38*, 176–182.

Aktaş, A. H., Dursun, Ş., Doğan, Ş., Kiyma, Z., Demirci, U., & Halıcı, İ. (2015). Effects of ewe live weight and age on reproductive performance, lamb growth, and survival in Central Anatolian Merino sheep. *Arc. Anim. Breed., 58*, 451-459.

Baeza-Rodríguez, J. J., Montaño-Bermúdez, M., Vega-Murillo, V. E., & Arechavaleta-Velasco, M. E. (2018). Linear and logistic models for multiple-breed genetic analysis of heifer fertility in Mexican Simmental–Simbrah beef cattle. *J. of Applied Animal Research, 46(1)*, 534-540.

Belsly, D. A., Kuh, E. & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential data and Source of Collinearity*. Wiley, New York.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics. 19(1)*, 15-18.

Copas, J. B. (1988). Binary regression model for contaminated data (with discuss). *Journal of the Royal Statistical Society, Series B., 50*, 225-265.

Estaghvirou, S. B. O., Ogutu, J. O. & Piepho, H.-P. (2014). Influence of Outliers on Accuracy Estimation in Genomic Prediction in Plant Breeding. *G3-Genes Genomes Genetics. 4*, 2317-2328.

Erdinç, S., Yeşilova, A., & Ser, G. (2017). Van gölü sahil şeridindeki zooplankton populasyon yoğunluğu değişiminin doğrusal olmayan regresyon yöntemleri kullanilarak incelenmesi. *Yüzüncü Yıl Üniversitesi Tarım Bilimleri Dergisi. 27 (1),* 58-64.

Eyduran, E., Özdemir, T., Çak, B., & Alarslan, E. (2005). Using of logistic regression in animal science. *Journal of Applied Sciences. 5(10)*, 1753-1756. doi:10.3923/jas.2005.1753.1756.

Fridendly, M., & Meyer, D. (2015). *Discrete Data Analysis with R Visualization and Modeling Techniques for Categorical and Count Data*. USA.

Gaskins, C. T., Snowder, G. D., Westman, M. K., & Evans, M. (2005). Influence of body weight, age and weight gain on fertility and prolificacy in four breeds of ewe lambs. *J. Anim. Sci., 83*, 1680-1689.

Gebre, T., Deneke, Y., & Begna, F. (2018). Seroprevalence and Associated Risk Factors of Peste Des Petits Ruminants (PPR) in Sheep and Goats in Four Districts of Bench Maji and Kafa Zones, South West Ethiopia. *Global Veterinaria, 20 (6)*, 260-270.

Habshah, M., Norazan, M. R., & Imon, A. H. M. R.(2009). The performance of diagnostic-robust Generalized Potentials for the identification of multiple High Leverage Points in Linear Regression. *Journal of Applied Statistics, 36*, 507-520.

Hadi, A. S. (1992). A new measure of overall potential influence in linear regression. *Computational Statistic Data Analysis, 14*, 1-27.

Hadi, A. S. & Simonoff, J. S. (1993). Procedure for the Identification of outliers in linear models. *J. Am. Stat. Asssoc., 88*, 1264-1272.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P.J., & Sathel, W. A.(1986). *Robust Statistics: The Approach Based on Influence Function*. Wiley, New York.

Heslot, N., Jannink, J. L. & Sorrells, M. E. (2013). Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. *Crop Sci., 53*, 921-933.

Hilbe, J. M. (2009). *Logistic Regression Models*. CRC Press, Newyork.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression. 2nd ed*. Wiley, New York.

Imon, A. H. M. R. (2006). Identification of High Leverage points in logistic regression. *Pak. J. Stat., 22*, 147-156.

Imon, A. H. M. R. & Hadi, A. S. (2008). Identification of multiple outliers in logistic regression. *Communications in Statistics Theory and Methods, 37*, 1697-1709.

Jennings, D. E. (1986). Outliers and residual distribution in logistic regression. *Journal of American Statistical Association, 81*, 987-990.

Korkmaz, M., Güney, S., & Yiğiter, Ş. Y. (2012). The importance of Logistic regression implementations in the Turkish livestock sector and logistic regression implementation/fields. *J. Agric. Fac. HR.U., 16(2)*, 25-36.

Midi, H. & Ariffin, S. B. (2013). Modified standardized pearson residual for the identification of outliers in logistic regression model. *Journal of Applied Sciences, 13*, 828-836.

Nurunnabi, A. A. M., Imon, R. A. H. M., & Nasser, M. (2010). Identification of multiple observations in logistic regression. *Journal of Applied Statistics, 37*, 1605-1624.

Nurunnabi, A., & Nasser, M. (2011). Outlier diagnostics in logistic regression: A supervised learning technique. *2009 International Conference on Medicine Learning and Computing IPCSIT, 3*, 90-95.

Nurunnabi, A. A. M. & West, G. (2012). *Outlier detection in logistic regression: A quest for reliable knowledge from predictive modeling and classification*. Paper presented at the 12th International Conference on Data Mining Workshops.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics, 9*, 705-724.

R-project (2018). The Comprehensive R Archive Network R for windows version 3.5.1.

Sanizah, A., Habshah, M., & Norazan, M. R. (2011). Diagnostic for residual outliers using deviance component in binary logistic regression. *World Applied Sciences Journal, 14 (8)*, 1125-1130.

Sarkar, S. K., Midi, H. & Rana, S. (2011). Detection of outliers and influeantial observations in binary logistic regression: An empirical study. *Journal of applied Statistics, 11(1)*, 26-35.

Takma, Ç., Güneri, Ö. İ., & Gevrekçi, Y. (2016). Investigation of Stillbirth rate using logistic regression analysis in Holstein Friesian calves. *Ege Ünv. Ziraat Fak. Derg., 53(3),* 245-250.

Via, S., Conte, G., Mason-Foley, C., & Mills, K. (2012). Localizing FSToutliers on a QTL map reveals evidence for large genomic regions of reduced gene exchange during speciation-with-gene-flow. *Mol. Ecol. 21*, 5546–5560.

Welsch, R. E. (1982). *Influence Functions and Regression Diagnostics*. Modern Data Analysis, New York Academic Press.

Yakubu A., Muhammed, M.M. & Musa-Azara, I.S. (2014). Application of multivariate logistic regression model to assess factors of importance influencing prevalence of abortion and stillbirth in Nigerian Goat Breeds. *Biotechnology Animal Husbandry, 30*,79-88.