International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 6 (Special Issue), 2007, pp 315-330

A New Graphical Display for Locating Multiple Influential Observations, High Leverage Points and Outliers in Linear Regression

A.H.M. Rahmatullah Imon
Dept. of Statistics
University of Rajshahi
Rajshahi-6205, Bangladesh
E-mail: imon_ru@yahoo.com

Paul Davies
Institute of Child Health
University of Birmingham
Edgbaston, Birmingham B15 2TT, U. K.
E-mail: pdaviesbch@blueyonder.co.uk

[Received April 2, 2005; Revised August 16, 2007; Accepted August 20, 2007]

Abstract

Numerous diagnostic plots are now available in the literature for identifying influential observations, high leverage points and outliers when using a linear regression model. Because of masking and swamping, most of the commonly used plots can produce misleading pictures when certain patterns of unusual cases are present in the data. A new diagnostic plot, named as the GP-DR plot, is proposed which is based on group deleted residuals and leverages so that masking and/or swamping do not affect it. Unlike the commonly used plots, this proposed plot retains the signs of the residuals which gives more insight into whether the outliers are influential or not.

Keywords and Phrases: Influential Observations, High Leverage Points, Outliers, Generalized Potentials, Deletion Residuals, L-R Plot, P-R Plot, GP-DR Plot.

AMS Classification: Primary 62 J20; Secondary 62 J05.

1 Introduction

Diagnostics methods are commonly used in all branches of regression analysis. A prime objective of these methods is the identification of unusual cases such as influential observations, high leverage points and outliers. Numerous diagnostic plots serving various purposes are now available in the statistical literature [see Atkinson (1985), Gray (1984), Hadi (1992), Ghosh (1996), Gray and Mayo (1997), Tsai et al. (1998), Montgomery et al. (2001), Chatterjee and Hadi (2006)]. In this paper we consider the class of plots that use residuals and leverages of any observation simultaneously to assess their influence on the fit. Although the identification of a single unusual case is often achieved satisfactorily by the use of single case deletion methods, these methods may be ineffective when a group of unusual cases have a combined effect to produce poor fitting of the model. Then methods based on group deletions are required. In section 2, we describe two types of diagnostic plots, the first of which is based on the least squares residuals and leverage values, and the second is based on single case deleted residuals and leverages. Then in section 3, we introduce a new type of diagnostic plot that involves group deleted residuals and leverages. The usefulness of this newly proposed diagnostic plot is demonstrated in section 4 by a variety of examples.

2 Leverage-Residual Plot

Consider a standard regression model

$$Y = X\beta + \in$$

where $Y$is an$n$-vector of observed responses, X is an $n \times k$ matrix representing $k$explanatory variables, β is a k-vector of unknown finite parameters and $\in$ is an $n-$vector of random disturbances with $E(\in) = 0$and $V(\in) = \sigma^2 I$. The traditionally used Ordinary Least Squares (OLS) estimator of β is $\hat{\beta} = (X^T X)^{-1} X^T Y$ and the vector of fitted values is $\hat{Y} = X\hat{\beta} = WY$. The matrix $W = X(X^T X)^{-1} X^T$ is often referred to as the weight or leverage matrix whose diagonal elements $w_{ii}$are termed leverages. The OLS residual vector $\hat{\in}$ is defined as $\hat{\in} = Y - \hat{Y}$. Observations corresponding to exceptionally large $\in$ values are termed outliers. Observations whose presence or absence can make a huge impact on the fitting of the model and hence the resulting analyses are called influential. High leverage points are those which exert too much influence in the X-space. Comprehensive reviews of different aspects of influential observations, high leverage points and outliers in linear regression are available in Rousseeuw and Leroy (1987), Chatterjee and Hadi (1988), Barnett and Lewis (1994) and Ryan (1997). A variety of diagnostic plots are now available in the statistical literature. In this section we consider the class of plots that use residuals and leverages to assess the influence of that observation on the fit. Gray (1984) proposed the Leverage-Residual (L-R) plot. In this diagram, the leverage value w_{ii} for each observation i, is plotted against the

square of a normalised form of its corresponding residual $\hat{\in}_i^2 \Big/ \sum_{i=1}^{n} \hat{\in}_i^2$. The bulk of the cases will be associated with low leverage and small residuals so that they cluster near the origin (0, 0). The unusual cases will have either high leverages or large residual components and so will tend to be separated from the bulk of the data. High leverage cases will be located in the upper area of the plot and observations with large residuals will be located in the area to the right.

Hadi (1992) commented that in the L-R plot the high leverage cases do not get sufficient emphasis in comparison with the cases having large residuals. He proposed an alternative plot, which he named the Potential-Residual (P-R) plot, where potentials are used as alternatives to the leverages. Writing the data matrix of k explanatory variables as $X = [x_1, x_2, \cdots, x_n]^T$, the i-th leverage value is defined as $w_{ii} = x_i^T (X^T X)^{-1} x_i$. The i-th potential is defined as

$$p_{ii} = x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i \tag{1}$$

where $X_{(i)}$ is the data matrix $X$with the i-th row deleted. It is difficult to obtain the theoretical distribution of $p_{ii}$and hence to get suitable threshold values for them to determine when they are large. A suitable cut-off point for p_{ii} could be

$$Mean\ (p_{ii}) + c.\ St.\ dev.\ (p_{ii}) \tag{2}$$

where$c$ is an appropriately chosen constant such as 2 or 3. This form is analogous to a confidence bound for a location parameter. But the problem with this threshold is that both mean and variance of p_{ii} may be sensitive to the presence of a single extreme value yielding a high cut-off point. To avoid this the mean and the standard deviation in (2) can be replaced by the median and the median absolute deviation (MAD) respectively.

In a P-R plot, the potential value $p_{ii} = w_{ii}/(1 - w_{ii})$for each observation i, is plotted against a normalized residual component

$$\frac{k\hat{\in}_i^2}{(1 - w_{ii}) \left(\sum \hat{\in}_i^2 - \hat{\in}_i^2\right)}.$$

Both the L-R plot and the P-R plot can be useful in assessing the influence of single observations. The P-R plot gives more emphasis to high leverage cases than the L-R plot. But masking and/or swamping effects can produce misleading plots for both and therefore diagnostic plots based on group deletions are needed.

3 Generalized Potentials and Deletion Residuals (GP-DR) Plot

If a group deletion method is able to produce a better set of residuals and leverages that are free of masking and swamping effects, then it can reasonably be expected to provide a graphical display also free from these effects. In this section, we define group deletion residuals and leverages. Denote a set of cases '*remaining*' in the analysis by R and a set of cases '*deleted*' by D. Also suppose that R contains $(n-d)$ cases after $d¡$ $(n-k)$ cases in D are deleted. Without loss of generality, assume that these observations are the last $d$rows of X and Y so that they can be partitioned as

$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix}$, $Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix}$

Then the weight matrix $W = X(X^TX)^{-1}X^T$becomes $W = \begin{bmatrix} U_R & V \\ V^T & U_D \end{bmatrix}$

where $U_R = X_R(X^TX)^{-1}X_R^T$ and $U_D = X_D(X^TX)^{-1}X_D^T$ are symmetric matrices of order $(n-d)$ and d respectively, and $V = X_R(X^TX)^{-1}X_D^T$ is an $(n-d)\times d$ matrix. Hence using the result of Henderson and Searle (1984), $(X_R^TX_R)^{-1}$ can be expressed as

$$(X_R^TX_R)^{-1} = (X^TX - X_D^TX_D)^{-1} = (X^TX)^{-1} + (X^TX)^{-1}X_D^T(I_D - U_D)^{-1}X_D(X^TX)^{-1} \tag{3}$$

where I_D is an identity matrix of order$d$. Then the vector of estimated parameters after the deletion of $d$observations, denoted by $\hat{\beta}^{(-D)}$, is obtained using (3) as

$$\hat{\beta}^{(-D)} = (\ X_R^TX_R\ )^{-1}\, X_R^TY_R = \hat{\beta} - (\ X^TX\ )^{-1}\, X_D^T\,(\ I_D - U_D\ )^{-1}\,\hat{\in}_D \tag{4}$$

where $\hat{\in}_D = Y_D - X_D\hat{\beta}$. Thus an $n\times 1$ vector of deletion residuals can be defined as

$$\hat{\in}^{(-D)} = Y - X\hat{\beta}^{(-D)} \tag{5}$$

From this the $j-th$ deletion residual is defined by

$$\hat{\in}_j^{(-D)} = Y_j - x_j^T\hat{\beta}^{(-D)} \quad , \ j = 1, 2, \cdots\cdots, n$$

Using (4), the deletion residual vector can be expressed in terms of the OLS residual vector $\hat{\in}$, and the vector of OLS residuals corresponding to the deleted points $\hat{\in}^{(-D)}$ as

$$\hat{\in}^{(-D)} = \ \hat{\in} + X\,(\ X^TX\ )^{-1}\, X_D^T\,(\ I_D - U_D\ )^{-1}\,\hat{\in}_D. \tag{6}$$

Deletion residuals are used extensively in regression diagnostics. Rousseeuw and Leroy (1987) proposed reweighted least squares (RLS) residuals where a full set of residuals

is estimated after the deletion of a group of suspect outliers. Here the outliers are identified by the least median of squares technique proposed by Rousseeuw (1984). RLS residuals have been widely used in robust regression. Many authors [see Hadi and Simonoff (1993) and Atkinson (1994)] also proposed similar group deletion residuals. Imon (1996) extended the definition of a single case deleted potential to group deletion which we now exploit to obtain a new graphical diagnostic by combining them with group deleted residuals. He named the resulting values as generalized potentials (GP) and studied their usefulness when a group of high leverage cases were masked. Define

$$w_{ii}^{(-D)} = x_i^T (X_R^T X_R)^{-1} x_i, \quad i = 1, 2, \cdots\cdots, n \tag{7}$$

so that $w_{ii}^{(-D)}$ is the i-th diagonal element of $X(X_R^T X_R)^{-1} X^T$ matrix. When $D = i$,we observe from (1) and (7) that

$$w_{ii}^{(-i)} = x_i^T \left(X_{(i)}^T X_{(i)} \right)^{-1} x_i = p_{ii} \ .$$

Suppose now that a further point i is removed from the remaining subset R and joins the deleted subset D. For any such i, it is easy to show that

$$w_{ii}^{-(D+i)} = \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \tag{8}$$

This tells us that the potential value of any case i, generated externally should be equivalent to the quantity $\frac{w_{ii}^{(-D)}}{1-w_{ii}^{(-D)}}$, when $w_{ii}^{(-D)}$ is generated internally on a reduced sample space R. The generalized potentials are defined for all members in a data set using (7) and (8) as

$$\begin{aligned} p_{ii}^{*} &= \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} \qquad for \ \ i \in R \\ &= w_{ii}^{(-D)} \qquad for \ \ i \in D \end{aligned} \tag{9}$$

where D is any arbitrary deleted set of points. Although the expression of generalized potentials is available for any arbitrary set of deleted cases, D, the choice of such a set is clearly important since the omission of this group determines the weights for the whole set.

For the selection of an appropriate D set, one could follow a stepwise method as suggested by Imon (2002). To form a set of candidate cases the j-th point of any regressor X_i is selected if it satisfies

$$X_{ij} \quad > Median\ (X_i) \quad \pm \quad c\ MAD\ (X_i) \tag{10}$$

where c is an appropriately chosen constant between 2 and 3. Different data points may satisfy rule (10) for each regressor. So initially all data points which satisfy rule (10) for any X_i are included as members of the deletion set. Then to see whether all members of the initial deletion set are points of relatively high leverages or not, one could consider p^*_{ii} to be large if above a threshold

$$p^*_{ii} \quad > Median\ (p^*_{ii}) + c\ MAD\ (p^*_{ii}). \tag{11}$$

If all members of the D set satisfy rule (11), then we can declare them as high leverage points. Since p^*_{ii}'s are measured in a similar scale, it may not matter much if a low leverage point is included in the deletion set. But in principle, it is better to put the low leverage cases back into the estimation subset R sequentially to re-compute p^*_{ii} values. This process is continued until all members of the deletion set individually satisfy rule (11) and the points thus identified are finally declared as high leverage points. It should be noted that a number of multivariate techniques are now available in the literature [see Peña and Prieto (2001), Maronna and Zamar (2002)], but we consider the stepwise method suggested by Imon (2002) because it is easy, computationally very simple and appears as efficient as the existing multivariate methods.

These results enable a simple graphical display of group deleted leverages and residuals. Generalized potentials are used as leverages and the robust RLS residuals as deletion residuals in 'generalized potentials - deletion residuals (GP-DR)' plot. Since the high leverage points need not be outliers and outliers may not be points of high leverage we may expect different deletion sets D from the computation of these two quantities. The main advantage of the GP-DR plot is that it is suitable for the data where masking and/or swamping make single case diagnostic plots misleading. This plot, unlike the L-R and P-R plots retains the signs of residuals, which can be very important when their interpretation. Another difference between the GP-DR plot and the other plots is that we do not propose the normalization of residuals or leverages. It is quite possible to suitably normalize generalized potentials and deletion residuals so that they could be measured on a similar scale, but for plots it is not crucial, as they are only scale factors. Since the bulk of the cases will be associated with low leverage and small residuals, most of the pairs $\left(\hat{\in}_i^{(-D)},\ p^*_{ii}\right)$ will cluster near the origin (0, 0). The unusual cases will have either high leverages or large residual components and will tend to be separated from the bulk of the cases. High leverage cases will be located in the upper area of the plot and observations with large residuals will be located either in the area to the left or to the right depending on their signs.

4 Examples

In this section we consider several well-known data sets that have been frequently used in the study of the identification of influential observations, high leverage points and outliers. Many data sets are now available in the literature for this purpose, but we

generally prefer to consider data sets where the situation generating the data is fully understood. Although simulated data sets are artificial in nature, they enable the comparison of different methods more reliably. Otherwise there is always uncertainty [see Cook and Hawkins (1990)] about which observations are actually unusual. For each of the examples considered we display three different plots; the L-R plot, the P-R plot and the GP-DR plot. For GP-DR plots we compute reweighted least squares residuals as deletion residuals (DR) by using the PROGRESS program of Rousseeuw and Leroy (1987) while the computation of generalized potentials (GP) are done with a simple program written in MINITAB. For these examples GP values are computed after the deletion of the cases which are above the threshold values as given by rule (11) of the previous section with $c = 2.5$.

The first example is the well-known Hawkins, Bradu, and Kass (1984) data. This artificial three-predictor data set contains 75 observations with 10 high leverage outliers (cases 1-10), 4 high leverage inliers (cases 11-14), and 61 low leverage inliers (cases 15-75). It has been reported [see Rousseeuw and Leroy (1987)] that all single case deletion methods not only fail to identify the true outliers but most of them also identify high leverage inliers as outliers. It has also been reported that [see Imon (2002)] that all commonly used leverage measures fail to focus on all of the high leverage points. Table 1 presents leverages, potentials, generalized potentials, OLS residuals and deletion residuals for this data. It should be noted that to compute GP the first 14 observations of this data set were omitted because the corresponding p^*_{ii} values for all of these 14 observations individually are over the threshold value which is 0.152. The RLS method identifies the first 10 observations as outliers and they are then omitted to compute the entire set of deletion residuals. Figures 1(a) to 1(c) show different diagnostic plots for this data. Both the L-R plot and the P-R plot fail to identify true outliers which cluster near (0,0) with 61 other clean observations. Only one of the 14 high leverage points, *i.e.* the observation no. 14 is identified as the point of high leverage and unfortunately 3 high leverage inliers (cases 11-13) are identified as outliers. However, the GP-DR plot is very successful in locating three groups of observations that are clearly separated from one another. All of the clean observations cluster around (0,0), 10 high leverage outliers (with positive signs) are located in the top right corner of the plot and 4 high leverage inliers are located in the top of the center of this plot.

The second example is the well-known stack loss data presented by Brownlee (1965) that has been extensively analyzed in the statistical literature. This three-predictor real data set (Air flow, Cooling water inlet temperature and Acid concentration) contains 21 observations with 4 outliers (observations 1, 3, 4, and 21) and 4 high leverage points (observations 1, 2, 3 and 21). Most of the commonly used diagnostic methods are able to identify only one (observation 21) out of 4 observations as outliers but they fail to identify even a single high leverage points. Hadi's potential method however swamps one good observation (case 17) as a high leverage point. Deletion residuals and generalized potentials for the stack loss data and are shown in Table 2 with the

OLS residuals, leverages and potentials. Observations 1, 3, 4 and 21 are identified as outliers by the RLS method and the rule based on generalized potentials described in the previous section mark observations 1, 2, 3 and 21 as high leverage points. Figure 2(a) shows that 3 outliers remain undetected when residual components are plotted against leverage values for this data. Moreover, all high leverage points are masked here. The P-R plot [see Figure 2(b)] also fails to identify 3 outliers and all the 4 high leverage points, but swamps a good observation (case 17) as a high leverage point. But the GP-DR method correctly identifies all outliers and high leverage points. In Figure 2(c), one high leverage point (case 2) is plotted on the extreme right, one low leverage positive outlier (case 4) is plotted on the top, two high leverage positive outliers (cases 1 and 3) are plotted on the top right and one high leverage negative outlier (case 21) is plotted on the bottom right corner of the graph. All these unusual observations are separated from the rest of the observations.

Finally we consider the artificial data set B of Peña and Yohai (1995). The main feature of this single predictor data set is that the two outlying observations (cases 9 and 10), which are also the points of equally high leverage, correspond to the true disturbances which are equal in magnitude but have opposite signs. The RLS and the GP are computed for this data set after the omission of these two high leverage outliers. The leverages, potentials, generalized potentials, OLS and deletion residuals for this data set are shown in Table 3 from which it can be seen that the OLS and deletion residuals for this data are almost identical. Because of a balancing effect, these two outliers do not cause any damage to the fitting of the model so that they are not jointly influential. These outliers have been termed [see Andrews and Pregibon (1978)] as outliers that do not matter. The GP-DR plot clearly focuses on the high leverage and the outlying behaviour of cases 9 and 10, but more crucially they exhibit the balancing effect of the two outliers. Observation 9 appears in the top right corner of this plot while observation 10 is located in the top left corner of the plot indicating that those are high leverage outliers but because of their balancing effects they are outliers that do not matter. But both of the L-R plot and the P-R plot fail to reveal this aspect as they reflect only the magnitude of the residuals not their signs.. This example emphasizes our point that diagnostic plots should retain the signs of the residuals for the better interpretation of the results.

5 Conclusions

In suggesting the GP-DR plot, it is anticipated that multiple case deletion in diagnostic plots will generally be preferable to single case deletion in giving reliable results. Our examples illustrate the different reasons why this may be so. Generalized potentials will give a clearer picture of influence effects than single case leverage values. We have chosen to illustrate the principle by using deletion residuals based on the reweighted least squares residuals but there are other alternatives which may be as good or superior to use in GP-DR. In particular the multiple case deletion residuals constructed on a conditional expectation principle suggested by Davies *et al.* (2004) appear to be better estimates of the true random disturbances than methods such as rewieghted least squares and the computational effort in getting them is not much greater. Though further numerical research to compare the performances of such methods is desirable to discover if there is an overall best, it is likely that the most crucial aspect is to avoid single case deletion methods and implement a reasonable multiple case method.

References

1. Andrews, D.F. and Pregibon, D. (1978). Finding the outliers that matter, *Journal of the Royal Statistical Society, Series-B*, 40, 85-93.

2. Atkinson, A.C. (1985). *Plots, Transformations, and Regression*, Clarendon Press, Oxford.

3. Atkinson, A.C. (1994). Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association*, 89, 1329-39.

4. Barnett, V. and Lewis, T.B. (1994). *Outliers in Statistical Data*, 3rd edition, Wiley, New York.

5. Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering*, 2nd edition, Wiley, New York.

6. Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression*, Wiley, New York.

7. Chatterjee, S. and Hadi, A.S. (2006). *Regression Analysis By Examples*, 4^{th} ed., Wiley, New York.

8. Cook, R.D. and Hawkins, D.M. (1990). Comment on 'Unmasking multivariate outliers and leverage points' by Rousseeuw, P.J. and van Zomeren, B.C., *Journal of the American Statistical Association*, 85, 640-44.

9. Davies, P., Imon, A.H.M.R. and Ali, M.M. (2004). A conditional expectation method for improved residual estimation and outlier identification in linear regression, *International Journal of Statistical Sciences*, 3 (Special volume in honour of Professor M.S. Haq), 191-208.

10. Ghosh, S. (1996). A new graphical tool to detect non-normality, *Journal of the Royal Statistical Society, Series-B*, 58, 691-702.

11. Gray, G.B. (1984). A simple graphic for assessing influence in regression, *Journal of Statistical Computation and Simulation*, 24, 121-34.

12. Gray, G.B. and Mayo, M.S. (1997). Graphical comparison of the resistant properties of regression estimators, *Proceedings of the section on Statistical Computing, American Statistical Association*, 142-47.

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

14. Hadi, A.S. and Simonoff, J.S. (1993). Procedure for the identification of outliers in linear models, *Journal of the American Statistical Association*, 88, 1264-72.

15. Hawkins, D.M., Bradu, D. and Kass, G.V. (1984). Location of several outliers in multiple regression data using elemental sets, *Technometrics*, 26, 197-208.

16. Henderson, H.V. and Searle, S.R. (1984). On deriving the inverse of a sum of matrices, SIAM *Review*, 22, 53-60.

17. Imon, A.H.M.R. (1996). *Subsample Methods in Regression Residual Prediction and Diagnostics*, Unpublished Ph.D. thesis, School of Mathematics and Statistics, University of Birmingham, U.K.

18. Imon, A.H.M.R. (2002). Identifying multiple high leverage points in linear regression, *Journal of Statistical Studies*, Special volume in honour of Professor Mir Masoom Ali, 207-18.

19. Maronna, R.A. and Zamar, R.H. (2002). Robust estimates of location and dispersion of high-dimensional data sets, *Technometrics*, 44, 307-17.

20. Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001). *An Introduction to Linear Regression Analysis*, 3^{rd}. ed., Wiley, New York.

21. Peña, D. and Prieto, F.J. (2001). Multivariate outlier detection and robust covariance estimation (with discussions), *Technometrics*, 43, 286-310.

22. Peña, D. and Yohai, V.J. (1995). The detection of influential subsets in linear regression by using an influence matrix, *Journal of the Royal Statistical Society, Series-B*, 57, 145-56.

23. Rousseeuw, P.J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871-80.

24. Rousseeuw, P.J. and Leroy, A. (1987). *Robust Regression and Outlier Detection*, Wiley, New York.

25. Ryan, T.P. (1997). *Modern Regression Methods*, Wiley, New York.

26. Tsai, C., Cai, Z. and Wu, X. (1998). The examination of residual plots, *Statistica Sinica,* 8, 445-65.