SORT 29 (1) January-June 2005, 1-10

Graphical display in outlier diagnostics; adequacy and robustness

Nethal K. Jajo*

University of Western Sydney

Abstract

Outlier robust diagnostics (graphically) using Robustly Studentized Robust Residuals (RSRR) and Partial Robustly Studentized Robust Residuals (PRSRR) are established. One problem with some robust residual plots is that the residuals retain information from certain predicated values (Velilla, 1998). The RSRR and PRSRR techniques are unaffected by this complication and as a result they provide more interpretable results.

MSC: 62-09, 62G35, 62J05, 62J20

Keywords: Masking; outlier; robust diagnostics; robust residuals; swamping.

1 Introduction

Graphical methods play an important role in fitting linear models in general and in outlier detection in particular. The informal graphical display and the formal testing procedures (numerical display) are complementary in outlier detection, and we emphasize that it should be implemented together. The informal graphical display is more useful than formal testing procedures since: first, the patterns of the residuals, the graphical displays dependant on them are often more informative than their magnitudes (Atkinson, 1985). Second, truth in the saying "one picture is worth thousand words". Third, a heavy computational requirement for numerical display is required, and in some cases the numerical display may fail to identify the potentiality of outlier observations.

* *Address for correspondence:* Dr. Jajo, N.K. University of Western Sydney, Blacktown Campus, Locked Bag 1797, Penrith South DC NSW 1797, Australia. Email: n.jajo@uws.edu.au

Received: October 2003
Accepted: June 2004

Several books devoting large portions to graphical presentations have been published, among them: Belsley, *et al.* (1980), Cook and Weisberg (1982), Atkinson (1985), Chatterjee and Hadi (1988), Barnett and Lewis (1994) and Hocking (1996).

Associating with others, we think robust residuals' plots are among the important graphical techniques in outlier detection. Many authors advocated robust residuals plots, among those are Rousseeuw and Leroy (1987). Different approaches are proposed by Velilla (1998), based on M-estimators using Huber's ψ function (Huber, 1973), GMM: three-step estimator of Simpson, *et al.* (1992) and GMS one-step estimator of Coakley and Hattmansperger (1993), who notices; "for high breakdown robust estimators, the residuals retain information on the regressor (here are X_i's) and this might complicates the interpretation of residuals plots". We are in agreement with others to a degree, but that is not the case when we are using RSRR or PRSRR plots in outlier detection.

The unaffected performance of using RSRR in outlier detection by the problem mentioned above will be shown through sections 2 – 4. Section 2 introduces RSRR and PRSRR, while Section 3 illustrates how to use RSRR /PRSRR plots and PRSRR probability plots in outlier detection. Real data examples will be given in Section 4, and Section 5 ends with a short conclusion.

2 Robust residuals

Consider the general linear regression model

$$Y_{n\times 1} = X_{n\times k}\beta_{k\times 1} + \epsilon_{n\times 1}, \tag{1}$$

where ϵ's have zero mean and variance matrix $\sigma^2 \times I_n$ (σ is unknown). The ordinary least squares (OLS) estimators $\hat{\beta}$ of β could be obtained by the minimization of

$$R(\hat{\beta}) = \sum \hat{r}_i^2 \quad \text{where} \quad \hat{r}_i = y_i - x_i^T\hat{\beta}. \tag{2}$$

While the robust estimators $\tilde{\beta}$ could be obtained by the minimization of

$$D(\tilde{\beta}) = \sum \rho(\tilde{r}_i/\sigma), \quad \tilde{r}_i = y_i - x_i^T\tilde{\beta}, \tag{3}$$

where ρ is a propriety function.

It's clear from equations 2 and 3, the robust estimators depend on the value of σ (while OLS does not). For this reason and other reasons given by Jajo (1999), we suggest using RSRR (robust residuals studentized by their robust scale estimators) and PRSRR for outlier detection. Moreover, Velilla's note (1998) is of relevance when we replace plotting robust residuals against regressors by: plotting RSRR against probability of each residuals (RSRR probability plots), plotting PRSRR against regressors (PRSRR plots) or plotting PRSRR against probability and regressors in three dimensions. This

mostly avoids (in fact is unaffected by) the problems of masking, swamping, or complicating the interpretation of residuals' plots, when residuals from high breakdown robust estimators retain information on the regressor variable, as mentioned by Velilla (1998) and Mckean *et al.* (1993, 1994).

2.1 RSRR

To obtain robust estimators; for the simple case of model 1, we use Theil-type method NMK; originally, the MK method was developed by Hussain and Sprent (1983) and Jajo (1999) proposed a modification termed NMK. Jajo's estimators are robust with high breakdown point (= 0.5) and can be defined as follows (Jajo, 1999):

$$\tilde{\beta} = \text{med}\{a_{i,i+m}\} \quad i = 1, 2, \ldots, m,\ m = n/2 \text{ if } n \text{ is even and } (n-1)/2 \text{ if } n \text{ is odd,}$$

where $a_{i,j} = (Y_j - Y_i)/(x_j - x_i)$, for $1 \le i < j \le n$, and $x_i \ne x_j$. The intercept parameter α can be estimated as $\tilde{\alpha} = \text{med}\{y_i - \tilde{\beta} x_i\}$.
In case of multiple linear regression (model 1) and to obtain robust estimators, we apply orthogonal modified Theil method based on NMK set of $a_{i,i+m}$ elements, for more details see Jajo and Wu (1998) and Jajo (1999) . For both cases, scaling robust residuals $\tilde{r}_i = y_i - x_i^T \tilde{\beta}$ by dividing them by corresponding s_i, we obtain RSRR as:

$$e_i = \tilde{r}_i / s_i, \tag{4}$$

where $s_i = s\sqrt{1 - p_{ii}}$. We recommend using s equal to the interquartile range (IR) since it achieves better results in robustness, as mentioned by Kianifard and Swallow (1989, 1996) in their comparison study of using variety of s as a robust estimate of the dispersion. For simple linear regression $p_{ii} = \frac{1}{n} + (x_i - \bar{x})^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2$ and $p_{ii} = x_i^T x_i \sum \lambda_r^{-1} \cos^2 \theta_{ir}$ for multiple case where λ_r is the r^{th} eigenvalue of $X^T X$ and θ_{ir} is the angle between x_i and the r^{th} normalized eigenvector of $X^T X$ (Chatterjee and Hadi 1988).

2.2 PRSRR

Suppose that

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

and

$$\tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \tilde{\beta}_2 x_{i2} + \cdots + \tilde{\beta}_k x_{ik},\ i = 1, 2, \ldots, n$$

are robust estimates for y_i. It follow that $\tilde{y}_i - \tilde{\beta}_\ell x_{i\ell}$ is a robust estimates of the i^{th} response

when all the predictors except the ℓ^{th} one ($x_{i\ell}$) are used. Hence the partial residuals at trial i can be defined as (du Toit *et al.* 1986):

$$\begin{aligned} p\tilde{r}_{i\ell} &= y_i - (\tilde{y}_i - \tilde{\beta}_{i\ell} x_{i\ell}), \qquad \ell = 1, 2, ..., k. \\ &= \tilde{r}_i + \tilde{\beta}_\ell x_{i\ell}. \end{aligned} \tag{5}$$

To obtain PRSRR $d_{i\ell}$ we scaled $p\tilde{r}_{i\ell}$ by the same way given in the preceding definition of RSRR.

3 Graphical display

An advantage of graphical display is that it can exhibits the effect of each observation, but some disadvantages occur in high breakdown robust residuals' plots as mentioned here by Velilla (1998). Suitability of a plot for a specified purpose could be the key to overcome the problem. Hadi (1993) states "the focus here is not on how the graph is constructed but rather on (a) what to graph, (b) what information can be extracted from a graph". Two-and three-dimensional graphs will be the main category of our graphical display. We used 2-D graphs as a way to compare the graphs of Velilla (1998) and Rousseeuw and Leroy (1987). We use 3-D graphs for explanation and confidence because software has made it possible to rotate 3-D plots and the choice of rotating position could be of interest.

3.1 RSRR Probability plot

In this plot, the probability (x-axis) axis is with $(i - .5)/n$ scale, where i (in 2-D) is the index of RSRR e_i, after ordering them in ascending magnitude. The y-axis is the RSRR e_i. In 3-D the i index is for x_i after ordering them also in ascending magnitude and represented as the x-axis; y-axis is the probability, and the z-axis is the e_i.

3.2 PRSRR and PRSRR Probability Plots

Plotting PRSRR as y-axis against corresponding x_i as x-axis will obtain PRSRR plot, while PRSRR probability plot is obtained by plotting PRSRR against corresponding x_i after ordering them in ascending magnitude. Probability axis is in the same scale as that in RSRR probability plot in 3-D.

For the two kind of plots, outliers will be recognized as the points that deviate markedly from the pattern of the whole observations. We emphasize that the graphical display is not enough for outlier detection, but it must be accompanied by the numerical display, which we do not present here to go along with Velilla's paper that is confined to the graphical display. For more details of using the two displays in outlier detection see Jajo (1999).

4 Real data examples

To illustrate the fact that the high ability of using robust residuals (RSRR or PRSRR) in outlier detection is unaffected by the retaining information on the regressor variable, possibly caused by the robust residuals, as Velilla (1998) notices. And to be nearer to Velilla's note, we make use of the two examples he used in his paper, and we add another well-known example for more explanation and confidence since it contains multiple outliers.

4.1 Gesell adaptive score data

These data are given by Mickey, Dunn and Clark (1967), and have been analyzed extensively in the statistical literature. These data contain 21 observations, with y regressed on x simply. Observation number 19 was regard as outlier. Figure 1: RSRR probability plot in 2-D marks observation 19 as outlier easily and so does RSRR probability in 3-D. Figure 3 of Velilla (1998) shows masking and swamping problems in RGMM-X plot and RGMM-INDEX plot, which confirms the improved achievement of our graphs.

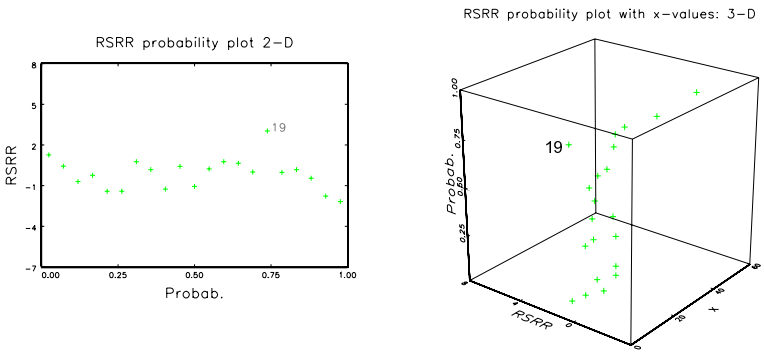

Figure 1: *RSRR probability plots for Gesell data.*

4.2 Salinity data

Ruppert and Carroll (1980) give this set of data, which contains 28 observations, three regressor variables and certainly two real outliers (observations number 5 and 16). Carroll and Ruppert (1985) and Atkinson (1985) perform residuals analysis for different models to fit salinity data and confirm certainly observations 5 and 16 as outliers. Rousseeuw and Leroy (1987) plot standardized LMS (Least Median of Squares) against estimated response and they mentioned that the horizontal band of points that must contain the data but not outliers is structureless. Velilla (1998) associated observation 16 as outlier, and he only recognized that case 5 has a large positive residual.

We agree that the salinity data is extremely complex, but still one of the better examples for masking, and there are a variety models that can be used to fit these data. Among other researchers, and for diagnostic purpose, we suggest our model to regress the water flow on bi-weekly average salinity, the salinity lagged two weeks and the trend. Our chosen model may not be the best possible one, but it is chosen because of its simplicity, going smoothly with our diagnostic and association with others in their results.

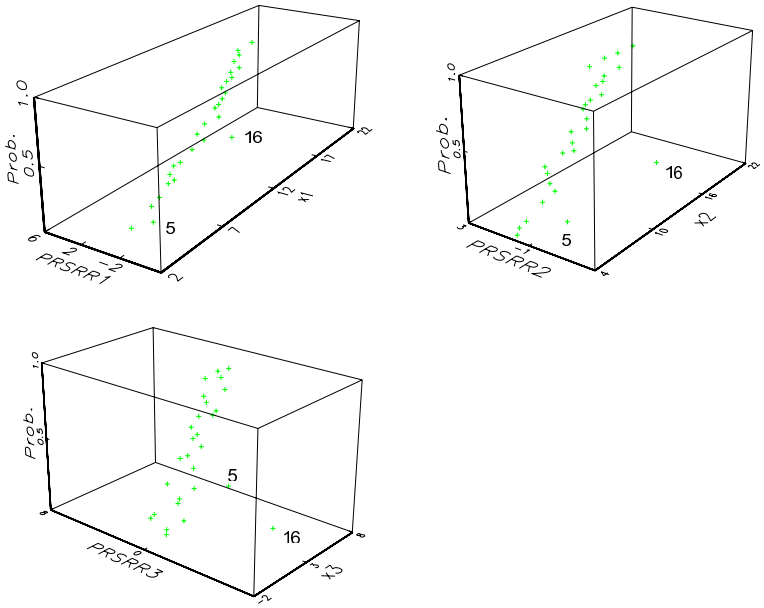

Figure 2: *PRSRR probability plots for Salinity data.*

From Figure 2, we can distinguish observations number 5 and 16 from others since it locate far from the main body of the observations, so regarded as outliers. The same could be mentioned for these observations in Figure 3. Both figures achieve better than Rousseeuw and Leroy (1987): Figure 4, page 84 and Velilla (1998): Figure 4 in detecting outliers without any problem of masking and swamping. Rousseeuw and Leroy (1987): Figure 4, shows swamping while Velilla (1998): figure 4, shows masking and swamping at plots of RGMM-INDEX, RGMM-X_1 and RGMM-X_2

4.3 Esoteric example (Dilemma data)

This set of data was used by Hocking and Pendleton (1983) and others. It contains 26 observations with 3 regressor variables, a constant term and 3 outliers (observations number 11, 17 and 18). Figure 4 uses PRSRR against corresponding regressor variables x_i. The three plots flag the outliers clearly without any problems. Moreover PRSRR probability plots in figure 5 show the same thing in 3-D. Using probability, the PRSRR

has been hung from PRSRR-x_i plane, and those outliers are located far from the shell of the other observations.

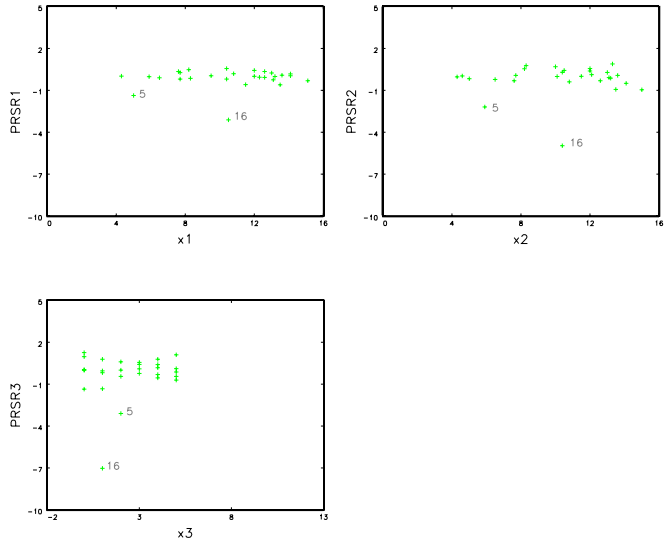

Figure 3: *PRSRR plots for Salinity data.*

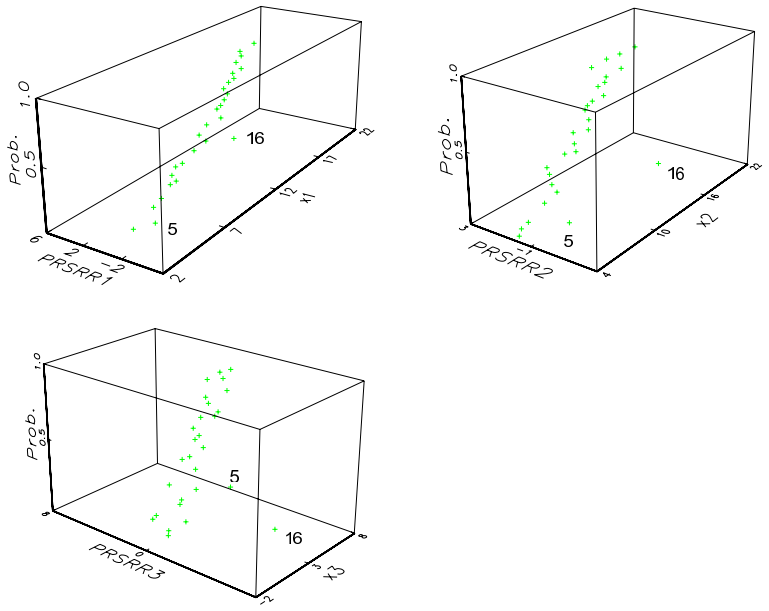

Figure 4: *PRSRR probability plots for esoteric data.*

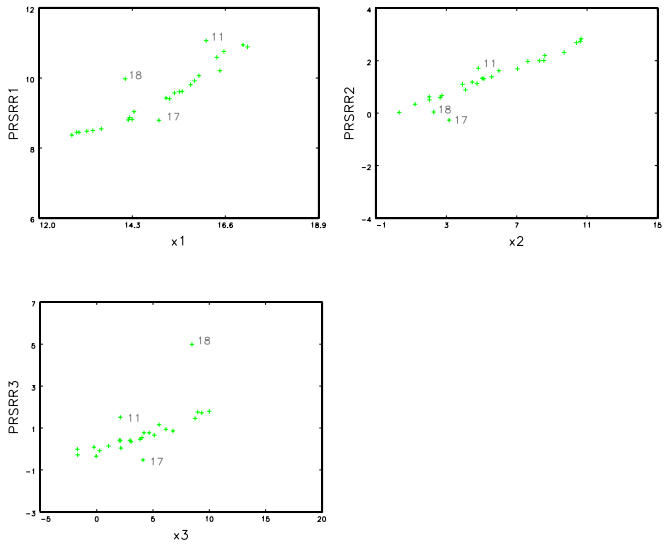

Figure 5: *PRSRR plots for esoteric data.*

5 Conclusions

Robust residuals might retain harmful information on regressor variables. But still those residuals play an important role in robust outlier diagnostics, specially when we are using RSRR or PRSRR through graphical or numerical display. The PRSRR and RSRR, PRSRR probability plots play an important role in outlier diagnostics, and their superiority in detecting multiple outliers is not affected by the enshroudment of outliers by each other, by other points, or by retaining harmful information on regressor variables that were caused by robust residuals in the residuals plots.

6 References

Atkinson, A.C. (1985). *Plots, Transformation, and Regression.* An introduction to graphical methods of diagnostic regression analysis. New York: Oxford University Press.

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

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: John Wiley.

Coakley, C. and Hettmansperger, T.P. (1993). A bounded influence, high break-down efficient regression estimator. *Journal American Statistical Association*, 88, 872-880.

Carroll, R.J. and Ruppert, D. (1985). Transformation in regression: a robust analysis. *Technometrics*, 27, 1-12.

Cook, R.D., and Weisberg, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

Cook, R.D., Hawkins, D.M. and Weisberg, S. (1992). Comparison of model misspecification diagnostics using residuals from least median of squares and least median of squares fits. *Journal American Statistical Association*, 87, 419-424.

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

du Toit, S.H.C., Steyn, A.G.W. and Stumpf, R.H. (1986). *Graphical Exploratory Data Analysis.* New York: Springer-Verlag.

Hocking, R.R (1996). *Method and Application of Linear Models.* New York: John Wiley & Sons. Inc.

Hocking, R.R. and Pendleton, O.J. (1983). The regression dilemma. *Communication in Statistics, part A-Theory and Methods*, 67, 388-394.

Huber, P. J.(1973). Robust Regression: asymptotic, conjectures, and Monte Carlo. *Annals of Statistics*, 1, 799-821.

Hussain, S.S. and Sprent, P. (1983). Non-Parametric Regression. *Journal of the Royal Statistical Society Series A. Part 2*, 146, 182-191.

Jajo, N. K. (1999). *Robust diagnostics of outliers in linear regression*, Ph.D. Thesis. Nanaki University, Tianjin, China.

Jajo, N.K. and Wu, X.Z. (1998). Robust diagnostics of outliers in multiple linear regression. *Proceedings of joint Statistical conference. Commemorating the 100th anniversary of Peking University.* Organized by the Institute of Mathematical Statistics and the Department of Probability and Statistics, Peking University, Peking, P.R.China.

Kianifard, F. and Swallow, W.H. (1989). Using recursive residuals, calculated on adaptive ordered observations, to identify outliers in linear regression. *Biometrics*, 45, 571-585.

Kianifard, F. and Swallow, W.H. (1996). A review of the development and application of recursive residuals in linear models. *Journal American Statistical Association*, 88, 1254-1263.

Mckean, J.W., Sheather, S.J. and Hettmansperger, T.P. (1993). The use and interpretation of residuals based on robust estimation. *Journal American Statistical Association*, 88, 1254-1263.

Mckean, J.W., Sheather, S.J. and Hettmansperger, T.P. (1994). Robust and high-breakdown fits of polynomial models. *Technometrics*, 36, 409-415.

Mickey, M.R., Dunn, O.J. and Clark, V. (1967). Note on use of stepwise regression in detecting outliers. *Computer & Biomed. Res.*, 1, 105-111.

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

Ruppert, D. and Carroll, R.J. (1980). Trimmed least squares estimation in the linear regression model. *Journal American Statistical Association*, 75, 828-838.

Simpson, D.G., Ruppert, D. and Carroll, R. (1992). On one-step GM estimates and stability of influences in linear regression. *Journal American Statistical Association*, 87, 439-450.

Swallow, H.W. and Kianifard, F. (1996). Using a robust scale estimates in detecting multiple outliers in linear regression. *Biometrics*, 52, , 545-556.

Velilla, S. (1998). A note on the behavior of residual plots in regression. *Statistics & Probability Letters*, 37, 269-278.