Working Paper 93-04
Statistics and Econometrics Series 04
February 1993

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

ON THE BEHAVIOUR OF RESIDUAL PLOTS IN ROBUST REGRESSION

Santiago Velilla*

Abstract

The behaviour of residual plots in robust regression might be distorted by the bias of the corresponding robust estimator. Examples of well-known data sets are discussed and reinterpreted.

Key Words

Bias of a Robust Estimator; Extreme Points; Least Median of Squares Regression; Outliers; Residual Plots.

*Departamento de Estadística y Econometría, Universidad Carlos III de Madrid. Research partially supported by DGICYT grant n° PB90-0266.

1. INTRODUCTION

The residuals obtained from a fit of a parametric statistical model are a popular building block for checking the adequacy of the assumptions underlying the fitted model and for detecting unusual patterns in the data. The paradigm of residual analysis is best illustrated in the context of a least squares fit of the multiple linear regression model, namely,

$$Y=X\beta+\varepsilon, \tag{1.1}$$

where Y is an nx1 vector of responses, X is an nxp known full rank matrix of regressors, β is a px1 vector of unknown parameters and ε is an nx1 vector of errors with $E[\varepsilon]=0$ and $V[\varepsilon]=\sigma^2 I_n$, where $\sigma^2>0$. If $\hat{\beta}=(X'X)^{-1}X'Y$ is the least squares estimator of β, the vector of fitted values is $\hat{Y}=X\hat{\beta}$ and the vector of least squares residuals is

$$e=Y-\hat{Y}=Y-X\hat{\beta}=(I_n-H)Y, \tag{1.2}$$

being $H=X(X'X)^{-1}X'$ the orthogonal projection matrix onto the linear manifold spanned by the columns of the design matrix X.

Residual analysis is generally conducted in a graphical way. The idea is to plot the residuals against another quantity orthogonal to them such that, under the null of a correctly specified parametric model, the expected behaviour of a plot contains no visible pattern. Observed patterns are then attributed to inappropriate assumptions. See, e.g., Cook and Weisberg (1982, chap. 2) for details and properties of residual plots. .

It is well known that the least squares estimator $\hat{\beta}$ is very sensitive to both outliers in Y and extremes in the rows of X and, therefore, several alternative robust estimators $\tilde{\beta}_n$ have been proposed. Plotting the residuals

$$\tilde{e}_i=y_i-x_i'\tilde{\beta}_n, \quad i=1, \ldots,n, \tag{1.3}$$

where x_i' is the ith row of X and y_i the ith response, is advocated by some authors, among others Rousseeuw and Leroy (1987, p. 92-3), as an after-fit diagnostic tool. Justification of the plot is not provided and, as a consequence, there is, some doubt about the interpretation and usefulness of a residual plot based in the nx1 vector $\tilde{e}=(\tilde{e}_i, i=1, \ldots, n)$. See, e.g., Cook and Weisberg (1990), or, recently, Cook, Hawkins and Weisberg (1992). See also Hettmansperger (1993) for an analysis of the behaviour of diagnostic plots in robust regression.

The aim of this paper is to present some results on the properties of residual plots in robust regression based on a large sample characterization of $\tilde{e}$. Section 2 establishes notation and presents some background for the results exposed in section 3. The theory is illustrated with examples of well-known data sets. Section 4 contains some final comments.

2. RESIDUALS AND ASYMPTOTIC BIAS OF A ROBUST ESTIMATOR

2.1 Robust estimators

We will consider estimators of the form

$$\tilde{\beta}_n=T_n(z_1, \ldots, z_n)=T[F_n], \tag{2.1}$$

where T[.] is a functional on a space of distributions, and F_n is the empirical distribution of the sample $z_i=(y_i,x_i')$, i=1, ...,n. Broadly speaking, an estimator is robust when it has "good" properties when the distribution of the pair (y,x') varies in the α-contamination neighborhood of Tukey

$$\mathcal{F}_\alpha=\{F: F=(1-\alpha)F_\theta+\alpha F^*\}, \tag{2.2}$$

where $\{F_\theta, \theta\in\Theta\}$ describes a central parametric model indexed by $\theta=(\beta',\sigma^2)'$, α is the fraction of contamination, $0\le\alpha\le.5$, and F^* is the distribution of the outliers which is unknown but, otherwise, arbitrary. The estimators considered are both consistent and Fisher consistent at the central model, i.e. for every θ, $\tilde{\beta}_n$ converges to β and $T[F_\theta]=\beta$. However, for general F in $\mathcal{F}_\alpha$, T[F] will be different from β and a suitable measure of the robustness of $\tilde{\beta}_n$ in large samples is the *maximum asymptotic bias* over $\mathcal{F}_\alpha$

$$B(T,F_\theta,\alpha)=\max\{b_M(T,F,\alpha):\ F\in\mathcal{F}_\alpha\},\tag{2.3}$$

where $b_M(T,F,\alpha)=\|T[F]-\beta\|_M$, $\|a\|_M=(a'Ma)^{1/2}$ and M is a pxp positive definite matrix. It is convenient to choose M in such a way that the maximum bias is invariant in front of affine transformations in the x_i.

Interest lies primarily in estimators with a finite-sample high breakdown point (BDP) as defined by Donoho and Huber (1983). For the model (1.1), the most studied high BDP estimators are the least median squares (LMS) $\tilde{\beta}_{LMS}$ of Rousseeuw (1984) which are defined as the solution of $\min_\beta$ median$(\varepsilon_1^2, \ldots,\varepsilon_n^2)$, where $\varepsilon_i=\varepsilon_i(\beta)=y_i-x_i'\beta$, i=1, ...,n. Other well-known robust estimators of β are Huber's (1973, p.800) M-estimators, several generalized M-estimators as in Krasker (1980) and Krasker and Welsch (1980), and S-estimators proposed by Rousseeuw and Yohai (1984). In what follows, the common notation $\tilde{\beta}_n$ will refer implicitly refer to any one of the estimators above.

Robust estimation is developed under the assumptions: (A.1) The sample $z_i=(y_i,x_i')'$, i=1, ...,n, is formed by independent and identically distributed observations from a pair (y,x')' with distribution F in F_α; and (A.2) The errors $\{\varepsilon_i\}$ are independent of the $\{x_i\}$. To fix ideas, we will assume that, under the central model, the distribution G of **x** is elliptically symmetric with $E_G[x]=0$ and $V_G[x]=\Sigma$ and that the distribution H of ε_i is $N(0,\sigma^2)$. Therefore, the central parametric model is formed by distributions of the form

$$F_\theta(y,x)=G(x)\Phi\left(\frac{y-x'\beta}{\sigma}\right),\tag{2.4}$$

where Φ is the N(0,1) distribution.

2.2 Residuals and bias of a robust estimator

The following relationship among the residual vector $\tilde{e}=Y-X\tilde{\beta}_n$, the vector of errors ε and the deviation $\tilde{\beta}_n-\beta$ holds

$$\tilde{e}=\varepsilon-X(\tilde{\beta}_n-\beta). \tag{2.5}$$

As it stands, expression (2.5) is not operational because of the untractability of the term $\tilde{\beta}_n-\beta$. To gain insight about the behaviour of $\tilde{e}$, we propose to replace the deviation $\tilde{\beta}_n-\beta$ by its asymptotic value $T[F]-\beta$ and consider

$$\tilde{e}_{as}=\varepsilon-X(T[F]-\beta)\simeq\tilde{e}. \tag{2.6}$$

We will assume that the sample size n is big enough so that the approximation (2.6) is meaningful. In agreement with (2.6), we introduce

$$\tilde{Y}_{as}=XT[F]\simeq\tilde{Y}, \tag{2.7}$$

where $\tilde{Y}=X\tilde{\beta}_n$ is the nx1 vector of fitted values corresponding to the robust estimator $\tilde{\beta}_n$. Notice that, in $\tilde{e}_{as}$, there are two independent sources of randomness, ε and X. This is in contrast with the usual least squares analysis where the only stochastic part is assumed to be ε.

3. RESIDUAL PLOTS UNDER CONTAMINATION

For any F in the parametric model $\{F_\theta, \theta\in\Theta\}$ Fisher's consistency yields $\tilde{e}_{as}=\varepsilon$ and, then, $E[\tilde{e}_{as}]=0$, $V[\tilde{e}_{as}]=\sigma^2 I_n$ and $cov(\tilde{e}_{as},\tilde{Y}_{as})=0$. Therefore, *if the central model is correct*, the residuals $\tilde{e}$ will likely reflect anomalies in ε. However, in robust regression it is assumed that the central model describes the data only approximately as expressed by the α-contamination neighborhood (2.2) above. We will explore the effect on a residual plot of three different types of contamination:

(*i*) Outliers in the x-direction which appear when x is generated by $(1-\alpha)G+\alpha G^*$. In the spirit of the neighborhood (2.2) above, the distribution F^* has the expression

$$F^*(y,x)=G^*(x)\Phi\left(\frac{y-x'\beta}{\sigma}\right). \tag{3.1}$$

(*ii*) Outliers in the y-direction associated with errors generated by $(1-\alpha)H+\alpha H^*$. The distribution F^* has the expression

$$F^{*}(y,x)=G(x)H^{*}\left(\frac{y-x'\beta}{\sigma}\right). \qquad (3.2)$$

(*iii*) A combination of (*i*) and (*ii*), namely a contamination of the form

$$F(y,x)=(1-\alpha)F_{\theta}(y,x)+\alpha_1 G^{*}(x)\Phi\left(\frac{y-x'\beta}{\sigma}\right)+\alpha_2 G(x)H^{*}\left(\frac{y-x'\beta}{\sigma}\right), \qquad (3.3)$$

$,\alpha=\alpha_1+\alpha_2$, $\alpha_i\geq 0$, $i=1,2$, i.e when $(1-\alpha)\times 100\%$ of the time we sample from the central model, $\alpha_1\times 100\%$ of the time from contaminated points in the x-space and $\alpha_2\times 100\%$ from aberrant errors.

To simplify matters we will assume that the distributions of ε and X have zero mean. The results which follow can be easily adapted to the case of general means. We illustrate the theory with examples. Since the method of least median squares is probably the most extended method in robust regression, the analysis of the examples below focuses on the residual vector $\tilde{e}=Y-X\tilde{\beta}_{LMS}$ and the corresponding vector of fitted values $\tilde{Y}=X\tilde{\beta}_{LMS}$. All computations have been made with the program PROGRESS of Rousseeuw and Leroy (1987). The purpose here is not to compare residual plots based on e with residual plots based on $\tilde{e}$ but merely to highlight some unexpected features which appear in residual plotting in robust regression.

3.1 Outliers in the x-direction. Under the contamination (3.1), we get

a) $E[\tilde{e}_{as}]=0$;

b) $V[\tilde{e}_{as}]=(a_1^2+a_2^2)I_n$, where

$a_1^2=\sigma^2+(T[F]-\beta)'\Sigma(T[F]-\beta)$;

$a_2^2=\alpha(T[F]-\beta)'(V_{G^*}[x]-\Sigma)(T[F]-\beta)$; and

c) $\mathrm{cov}(\tilde{e}_{as},\tilde{Y}_{as})=(b_1+b_2)I_n$, where

$b_1=-(T[F]-\beta)'\Sigma T[F]$;

$b_2=-\alpha(T[F]-\beta)'(V_{G^*}[x]-\Sigma)T[F]$.

These results highlight three potential sources of concern while using a plot based on the components of the vector $\tilde{e}$: *i*) From b) the variance of the residuals is "inflated" by a term which depends on the bias $T[F]-\beta$ of the estimator. It is known that an estimator can have high BDP≅.5 and, at the same time, high $B(T,F_\theta,\alpha)$. Therefore, a plot based on $\tilde{e}$ might be too erratic to allow for a proper interpretation; *ii*) Also from b), the bias $T[F]-\beta$ and the matrix $V_G{*}[x]$ can combine together to produce a large variance. Therefore, a case might be pinpointed in a plot simply because corresponds to an extreme row of X and not to an outlying response; *iii*) From c), the plot of the pairs $(\tilde{e}_i,\tilde{Y}_i)$, $i=1, \ldots,n$, can contain a linear trend as expressed by the nonnull correlation coefficient of the pairs $(\tilde{e}_{as,i},\tilde{Y}_{as,i})$, $i=1, \ldots,n$.

EXAMPLE 3.1 Salinity data. The data and previous analysis of them are described in page 82 and ff. of Rousseeuw and Leroy (1987). The model contains n=28 cases and p=3 regressors plus an intercept. It seems to be common agreement about the fact that cases 3, 5 and 16 are extreme rows of the matrix X. Figure 1.a) is an index plot of the raw lms residuals. The plot draws attention towards points 5 an 16, which are not outliers but extreme rows in X, and to points 23 and 24 which are good data points. A possible explanation of this misleading behaviour of the plot could be found in property b) above. Figure 1.b) is a plot of the raw lms residuals versus the fitted lms responses where, as expected from c) above, there is a linear trend as remarked by the superimposed least squares straight line.

Figure 1. Salinity data. Residual plots in lms robust regresion.

EXAMPLE 3.2 Artificial data set. These data appear in page 37 of Rousseeuw and Leroy (1987) and are used as an illustration of the use of the program PROGRESS. The model is simple linear regression with n=20 cases. Case 6 is an extreme row which appears clearly pinpointed in the index plot of the raw lms residuals displayed in figure 2.a). Figure 2.b) is the plot of the lms residuals versus the lms fitted response.

Figure 2. Artificial data. Residual plots in lms robust regression

3.2 Outliers in the y-direction. Under the contamination (3.2), we get

d) $E[\tilde{e}_{as}]=0$;

e) $V[\tilde{e}_{as}]=(a_1^2+a_3^2)I_n$, where

$$a_3^2=\alpha(\mathrm{var}_{H^*}[\varepsilon_1]-\sigma^2);$$

f) $\mathrm{cov}(\tilde{e}_{as},\tilde{Y}_{as})=b_1I_n$.

d), e) and f) can be interpreted as in *i*) and *iii*) above.

3.3 Mixed contamination scheme. Results and consequent interpretations can be adapted for a mixed contamination scheme of the form (3.3). The following example illustrates the theory.

EXAMPLE 3.3 Brownlee's stackloss data. This example refers to a model with n=21 points and p=3 regressors. Description of the data, meaning of the variables and mentions to previous analysis of this data set can be seen in Rousseeuw and Leroy (1987, p. 76). There seems to be a general agreement that cases 1, 3, 4 and 21 are anomalous. Case 2 is also suspicious for some authors. Cases 1 and 2 are extremes, while cases 3,4 and 21 are outliers. Figure 3.a) is an index plot of the raw lms residuals which pinpoints correctly cases 3, 4 and 21 and, misleadingly, cases 1 and 2. The plot of the raw lms residuals versus the fitted lms responses, as shown in figure 3.b), contains again a linear trend.

Figure 3. Stackloss data. Residual plots in lms robust regresion.

4. DISCUSSION

Residual plots in linear least squares regression are simple to use and the information conveyed by these plots is, in general, powerful and elegant. These two properties might be responsible of having extended among practitioners the idea that a plot of residuals is a universal diagnostic tool even in contexts different from a least squares fit of the model (1.1). The justification of the procedure has been, in general, considered unnecessary.

The examples analyzed in this paper show that a residual plot in LMS robust regression can produce a misleading impression due, among other reasons, to the unexpected contribution of the bias of the estimator. Although the explanation provided is asymptotic in nature, as is most of the work and the results in robust regression, it seems to be a reasonable agreement between the consequences of (2.6) and examples 3.1, 3.2 and 3.3. Characterization of the plots in small samples remains as an open problem. In this situation, there are, however, some additional warnings regarding the finite sample behaviour of high breakdown robust estimators. The anomalies are associated with the so called exact-fit property. See Stefanski (1991) and Hettmansperger and Sheather (1992).

REFERENCES

COOK, R.D. & WEISBERG, S. (1982). *Residuals and Influence in Regression*. New York and London: Chapman and Hall.

COOK, R.D. & WEISBERG, S. (1990). Regression diagnostics: Robust *versus* least squares residuals. University of Minnesota, School of Statistics, Dept. of Applied Statistics, Tech. Report No. 538

COOK, R.D., HAWKINS, D. & WEISBERG, S. (1992). Comparison of model misspecification diagnostics using residuals from least mean of squares and least median of squares fits. *J. Am. Statistic. Ass.*, **87**, 419-424.

DONOHO, D.L. & HUBER, P,J. (1983). The notion of breakdown point. In: P. Bickel, K. Doksum and J.L. Hodges Jr., eds, *A Festschrift for Erich L Lehmann*. Belmont, CA: Wadsworth.

HETTMANSPERGER, T.P. & SHEATHER, S.J. (1992). A cautionary note on the method of least median squares. *The American Statistician*, **46**, 79-83.

HETTMANSPERGER, T.P. (1993). Robust fittings and diagnostic plots. To appear in the *Journal of Nonparametrics*.

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, **1**, 799-821.

KRASKER, W.S. (1980). Estimation in linear regression models with disparate data points. *Econometrica*, **48**, 1333-1346.

KRASKER, W.S. & WELSCH, R.E. (1982). Efficient bounded-influence regression estimation. *J. Am. Statistic. Ass.*, **77**, 595-604.

ROUSSEEUW, P.J. (1984). Least median of squares regression. *J. Am. Statistic. Ass.*, **79**, 871-880.

ROUSSEEUW, P.J. & YOHAI, V. (1984). Robust regression by means of S-estimators. In: Franke, Handle and Martin, eds, *Robust and Nonlinear Time Series Lecture Notes in Statistics., No.* 26. New York: Springer.

ROUSSEEUW, P.J. & LEROY, A.M. (1987). *Robust Regression and Outlier Detection*. New York: J. Wiley.

STEFANSKI, L.A. (1991). A note on high-breakdown estimators. *Statistics and Probability Letters*, 11, 353-358.

FIGURES AND CAPTIONS

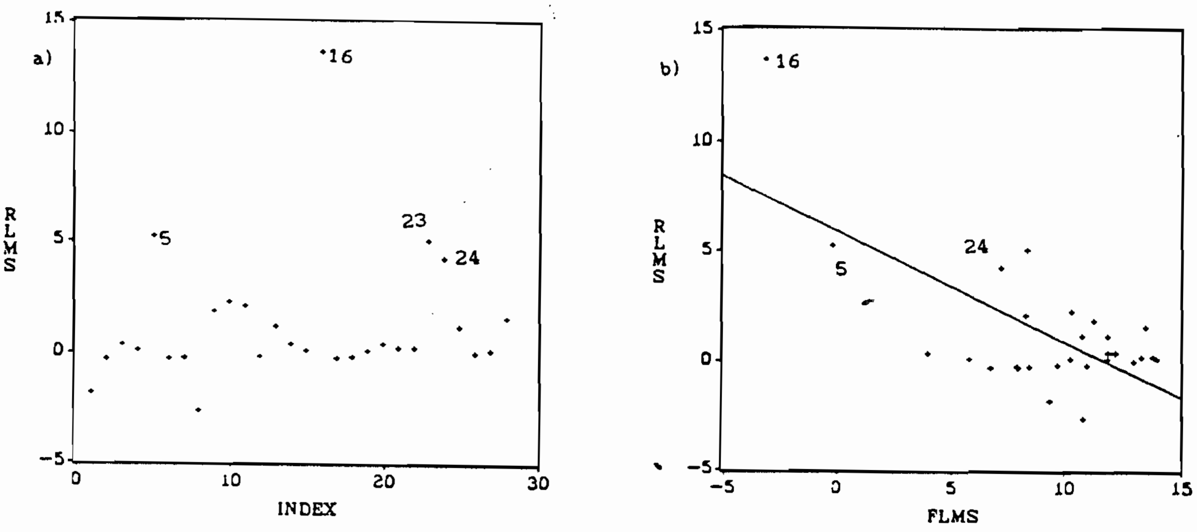

Figure 1. Salinity data. Residual plots in lms robust regresion.

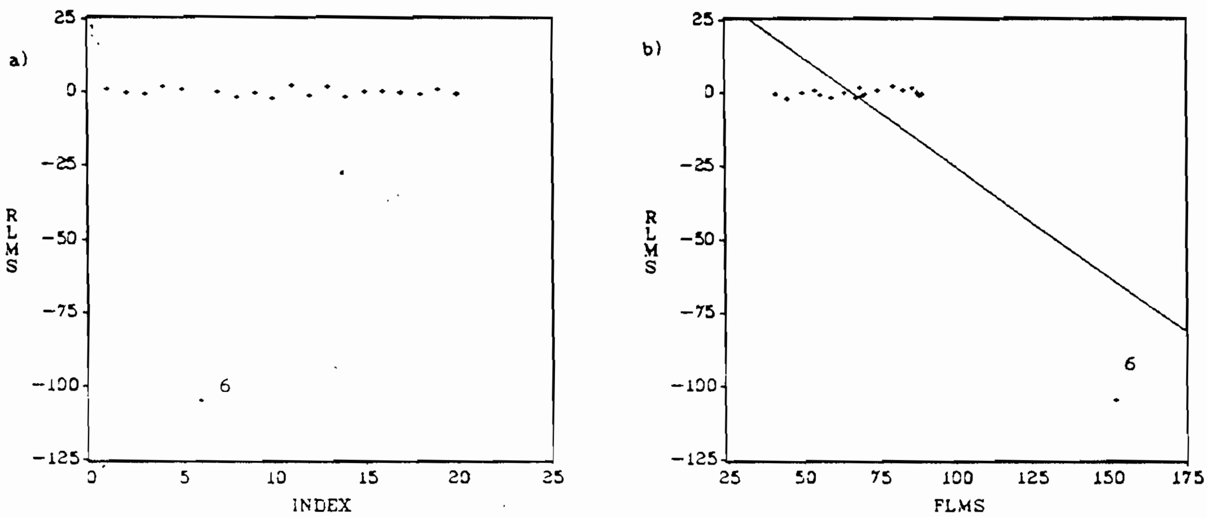

Figure 2. Artificial data. Residual plots in lms robust regression

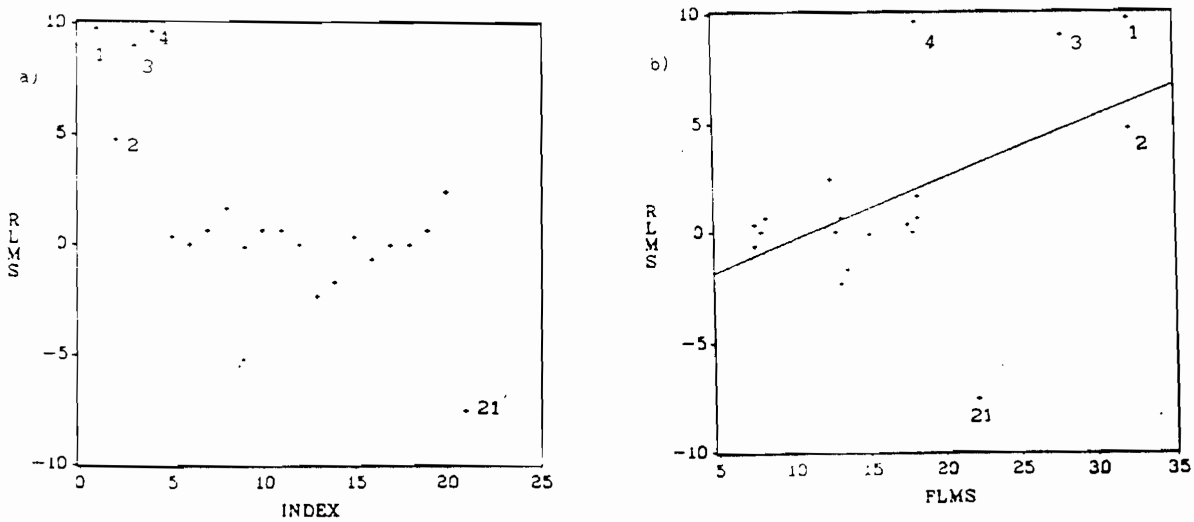

Figure 3. Stackloss data. Residual plots in lms robust regresion.