10.48047/jocaaa.2026.35.06.31

Six stage Identification and Assessment of Outliers and Influential Points in Regression Models

Shaik Hussan Saheb[1], B Sarojamma[2]

[1]Research scholar, Department of Statistics, S.V University, Tirupati.
[2]Professor and HoD, Department of Statistics, S.V University, Tirupati.
Corresponding author: saroja14397@gmail.com

Abstract

In regression analysis, a few unusual observations can change the fitted model. Such observations are of three kinds: outliers, which have a response far from the fitted line; high-leverage points, which are far from the other observations in the predictor space; and influential points, which change the estimated coefficients when they are removed. These three ideas are related but not the same, and using one measure alone can give a wrong picture. This paper explains, with simple mathematics, how the residual, the leverage and the influence of an observation are connected, and shows why the common diagnostic measures such as studentized residuals, leverage values, Cook's distance, DFFITS, DFBETAS and the covariance ratio can each fail when used on their own. It also explains why groups of unusual observations can hide one another (masking) or make normal observations look unusual (swamping). A six-stage procedure is proposed that combines residual, leverage and influence information, classifies each observation, and compares the model with and without the suspicious observations. Finally, simple guidance is given on when an observation should be kept, checked, down-weighted or removed. The paper stresses that an influential observation is not necessarily a wrong observation.

Keywords

Regression diagnostics; outliers; influential observations; leverage; hat matrix; Cook's distance; masking and swamping; robust regression.

1. Introduction

The method of least squares remains the workhorse of applied regression analysis because it is simple to compute, has an elegant geometric interpretation and is optimal among linear unbiased estimators when the classical assumptions hold. The same properties that make it attractive also make it fragile. Because the fitted hyperplane minimises the sum of squared vertical distances, a single observation that lies far from the bulk of the data can pull the fit towards itself, and the extent of this pull depends not only on how far the observation lies from

10.48047/jocaaa.2026.35.06.31

the fitted surface but also on where it sits in the space of the predictors. A data analyst who examines only the residuals will miss observations that have been accommodated by the fit precisely because they exert a strong pull; a data analyst who examines only the position of the observations in predictor space will flag points that, although remote, are perfectly consistent with the regression relationship and in fact improve its precision.

The vocabulary of regression diagnostics reflects these distinctions. An observation whose response is unusual given its predictor values is called a vertical outlier. An observation whose predictor values are remote from the centre of the predictor cloud is called a high-leverage point. An observation whose removal substantially changes the estimated coefficients, the fitted values, the residual variance or the precision of the estimates is called an influential observation. A high-leverage point whose response agrees with the regression relationship is often called a good leverage point, and one whose response disagrees with the relationship is called a bad leverage point. When several atypical cases occur together, two further phenomena arise: masking, in which the presence of one atypical case hides another so that neither is detected, and swamping, in which atypical cases distort the fit so badly that ordinary observations appear anomalous.

Despite a large body of literature, these concepts are still frequently confused in applied work, and diagnostics are still frequently reported as a list of numbers with fixed cut-offs and no interpretation. The purpose of this paper is threefold. First, we set out the mathematical structure that links residual size, leverage and deletion influence, and use it to show why no single measure is adequate. Second, we propose a staged procedure in which residual, leverage and influence information are combined into a joint classification of each observation, with explicit attention to the group-deletion problem that underlies masking and swamping. Third, we translate the resulting classification into a decision protocol for the analyst, emphasising that an influential observation is not automatically an erroneous one.

The remainder of the paper is organised as follows. Section 2 reviews the literature up to the end of the last decade and identifies the gap addressed here. Section 3 introduces the model, notation and the classical diagnostic quantities, and describes the proposed six-stage methodology. Section 4 presents the main research: the algebraic decompositions, the failure modes of the individual measures, the mechanism of masking and swamping, the joint classification rule and the sensitivity comparison. Section 5 concludes.

2. Review of Literature

10.48047/jocaaa.2026.35.06.31

2.1 Origins: residuals, leverage and the hat matrix

The systematic study of atypical observations in regression grew out of two older traditions: the testing of a single outlier in a sample, surveyed comprehensively by Barnett and Lewis (1994) and Hawkins (1980), and the study of residuals as a tool for model checking (Anscombe and Tukey, 1963; Draper and Smith, 1966). Tests for a single outlier in simple linear regression were given by Tietjen, Moore and Beckman (1973) and extended to the general linear model by Ellenberg (1976), whose test statistic is the maximum absolute externally studentized residual. Beckman and Cook (1983) reviewed this literature and clarified the distinction between an outlier as a discordant value and an outlier as a contaminant generated by a different mechanism.

The recognition that residuals alone are insufficient came with the analysis of the projection (hat) matrix. Hoaglin and Welsch (1978) gave a detailed account of the diagonal elements of the hat matrix, their interpretation as leverage, their bounds and their average value, and popularised the rule that a leverage value exceeding twice the average deserves attention. The key fact, that the variance of a residual is proportional to one minus the leverage, implies that observations of high leverage tend to have small residuals no matter how discordant their responses, so that raw residuals are systematically uninformative exactly where they are most needed. Behnken and Draper (1972) had already noted the uneven variance of residuals; Hoaglin and Welsch connected this to the geometry of the predictor space.

2.2 Deletion diagnostics and influence

Cook (1977) introduced the distance that now bears his name, which measures the displacement of the coefficient vector when a single case is deleted, scaled by the estimated covariance matrix of the coefficients. Cook (1979) extended the idea to the influence of a case on subsets of coefficients and on fitted values, and emphasised that influence is a joint property of residual and leverage. Belsley, Kuh and Welsch (1980) proposed a family of related deletion statistics, including DFFITS (the scaled change in the fitted value of a case when it is deleted), DFBETAS (the scaled change in each individual coefficient) and the covariance ratio (the change in the generalised variance of the estimates). They also provided the size-adjusted cut-offs that remain in common use. Draper and John (1981) compared these measures and proposed additional influence statistics based on the residual sum of squares. Velleman and Welsch (1981) showed that all the single-case statistics can be computed from one least-squares fit without refitting, which made routine diagnostic screening practical.

10.48047/jocaaa.2026.35.06.31

Cook and Weisberg (1982) gave the first monograph-length treatment of residuals and influence, unifying the deletion statistics through the influence curve of Hampel (1974). Atkinson (1985) added graphical methods and half-normal plots of diagnostic quantities. Chatterjee and Hadi (1986), in an influential review, organised the literature around the three concepts of outlier, high-leverage point and influential observation, showed by example that each can occur without the others, and warned against the mechanical use of cut-offs. Their monograph (Chatterjee and Hadi, 1988) developed the algebra of case deletion in detail and introduced a number of further measures.

A different perspective was opened by Cook (1986), who replaced the deletion of a case by a small perturbation of the model or the data and examined the curvature of the likelihood displacement in the direction of the perturbation. This local-influence approach avoids the discreteness of deletion and can be applied to case weights, response perturbation and predictor perturbation. Lawrance (1988) and Wu and Luo (1993) developed the local approach further, and it has since been extended to many model classes.

2.3 Multiple outliers, masking and swamping

The limitations of single-case deletion when several atypical cases are present were recognised early. Andrews and Pregibon (1978) proposed a statistic for subsets of observations based on the change in the volume of the data ellipsoid. Cook and Weisberg (1982) discussed the joint influence of pairs of cases. Atkinson (1986) demonstrated masking in a series of worked examples and showed how a preliminary robust fit can unmask hidden outliers. Rousseeuw and Leroy (1987) devoted a substantial part of their monograph to the failure of classical diagnostics in the presence of grouped outliers and leverage points. Hadi and Simonoff (1993) proposed a forward procedure that starts from a clean basic subset and adds cases one at a time, testing each addition, and showed that it resists masking. Davies and Gather (1993) gave a formal treatment of outlier regions and the concept of masking and swamping breakdown points. Peña and Yohai (1995) introduced an influence matrix whose eigenvectors reveal groups of jointly influential cases, and Peña (2005) later proposed a statistic that measures the influence of the whole sample on the fitted value of each case, which is sensitive to groups of outliers that single-case measures miss. Lawrence (1995) analysed the relationship between deletion influence and masking in detail.

10.48047/jocaaa.2026.35.06.31

2.4 Robust regression and high-breakdown diagnostics

A parallel line of work seeks estimators that are not unduly affected by atypical cases, so that residuals from the robust fit can be used to identify them. Huber (1973, 1981) introduced M-estimators, which bound the influence of large residuals but remain vulnerable to leverage points. Hampel, Ronchetti, Rousseeuw and Stahel (1986) systematised the influence-function approach. Rousseeuw (1984) proposed least median of squares and least trimmed squares, the first estimators with the highest possible breakdown point; Rousseeuw and Yohai (1984) introduced S-estimators, and Yohai (1987) combined high breakdown with high efficiency in the MM-estimator. Rousseeuw and van Zomeren (1990) proposed a diagnostic plot of robust standardised residuals against robust distances in predictor space, which classifies each observation as regular, vertical outlier, good leverage point or bad leverage point; this plot is the direct ancestor of the classification rule developed in Section 4. Hadi (1992) proposed a single overall measure of potential influence. Atkinson and Riani (2000) developed the forward search, which orders the observations by closeness to a robust fit and monitors diagnostic quantities as the subset grows, so that masked outliers enter last and are revealed by sharp changes in the monitored statistics. Imon (2005) proposed generalised versions of the classical deletion diagnostics computed from a group-deleted fit, and Habshah, Norazan and Imon (2009) proposed diagnostic-robust generalised potentials for detecting multiple high-leverage points.

2.5 Reviews and textbooks

Comprehensive treatments of regression diagnostics are provided by Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982), Atkinson (1985), Chatterjee and Hadi (1988), Fox (1991), Rousseeuw and Leroy (1987) and, in the robust context, Maronna, Martin and Yohai (2006). Textbook expositions with practical emphasis include Weisberg (1985), Draper and Smith (1998), Myers (1990), Montgomery, Peck and Vining (2012) and Kutner, Nachtsheim and Neter (2004). Kianifard and Swallow (1996) reviewed the recursive-residual approach to diagnostics. Hubert, Rousseeuw and Van Aelst (2008) surveyed high-breakdown multivariate methods, including their use in regression diagnostics.

2.6 Research gap

Three gaps emerge from this review. First, although the distinction between outlier, leverage and influence has been recognised since Chatterjee and Hadi (1986), the algebraic reasons why each classical measure fails, and the exact conditions under which it does so, are scattered across the literature and are rarely presented in one place with a common notation. Second, the

10.48047/jocaaa.2026.35.06.31

deletion statistics and the robust classification plot of Rousseeuw and van Zomeren (1990) are usually treated as alternatives rather than as complementary evidence to be combined. Third, the step from diagnosis to decision—retain, investigate, down-weight or remove—receives little formal attention. This paper addresses all three by deriving the joint structure of residual, leverage and influence within a single framework, by proposing a staged procedure that combines them, and by stating an explicit decision protocol.

3. Methodology

3.1 Model and notation

Consider the classical linear regression model

$$y = X\beta + \varepsilon, \quad \varepsilon \sim (0, \sigma^2 I_n) \tag{1}$$

where y is the $n \times 1$ vector of responses, X is the $n \times p$ matrix of predictors (including a column of ones for the intercept) of full column rank, β is the $p \times 1$ vector of unknown coefficients and ε is the vector of errors with mean zero and constant variance σ^2. The i-th row of X is written x_i^T. The ordinary least squares (OLS) estimator and its covariance matrix are

$$\beta = (X^T X)^{-1} X^T y, \quad Var\,\beta = \sigma^2 (X^T X)^{-1}. \tag{2}$$

The vector of fitted values is $\hat{y} = Hy$, where

$$H = X(X^T X)^{-1} X^T \tag{3}$$

is the hat matrix, a symmetric idempotent projection of rank p. The residual vector is $e = y - \hat{y} = (I - H)y$, the residual sum of squares is $RSS = e^T e$, and the usual estimate of the error variance is $s^2 = RSS/(n - p)$.

3.2 Leverage

The diagonal elements h_{ii} of H are the leverage values,

$$h_{ii} = x_i^T (X^T X)^{-1} x_i, \quad 0 \le h_{ii} \le 1, \quad \sum_{i=1}^{n} h_{ii} = p. \tag{4}$$

Because $\partial \hat{y}_i / \partial y_i = h_{ii}$, the leverage measures the extent to which the i-th fitted value is determined by its own response. When the model contains an intercept, h_{ii} is an increasing function of the Mahalanobis distance of x_i from the centroid of the predictors: with $\bar{x}$ the vector of predictor means (excluding the intercept) and S the sample covariance matrix of the predictors,

$$h_{ii} = \frac{1}{n} + \frac{1}{n-1} (x_i - \bar{x})^T S^{-1} (x_i - \bar{x}). \tag{5}$$

10.48047/jocaaa.2026.35.06.31

Thus leverage is purely a property of the predictor configuration and is independent of the response. The conventional screening rule flags $h_{ii} > 2p/n$ (Hoaglin and Welsch, 1978), with $3p/n$ sometimes used when p is small relative to n.

3.3 Residuals

Under the model, $Var(e) = \sigma^2(I - H)$, so that $Var(e_i) = \sigma^2(1 - h_{ii})$. The internally studentized residual scales each residual by its own estimated standard deviation,

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}, \tag{6}$$

and satisfies $r_i^2 \leq n - p$. The externally studentized (or deletion) residual uses the variance estimate $s_{(i)}^2$ obtained after deleting case i,

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}} = r_i\sqrt{\frac{n - p - 1}{n - p - r_i^2}}, \tag{7}$$

and under normality follows a Student t distribution with $n - p - 1$ degrees of freedom. The externally studentized residual is preferred for outlier testing because the variance estimate is not inflated by the case under scrutiny; a Bonferroni-adjusted comparison with $t_{n-p-1}(\alpha/2n)$ controls the family-wise error rate of the maximum absolute value (Cook and Weisberg, 1982). The PRESS (prediction) residual is the error in predicting y_i from a fit that excludes case i,

$$e_{(i)} = y_i - x_i^T\beta_{(i)} = \frac{e_i}{1 - h_{ii}}, \quad PRESS = \sum_{i=1}^{n} e_{(i)}^2. \tag{8}$$

3.4 Single-case deletion and influence measures

Let $\beta_{(i)}$ denote the OLS estimate computed without case i. The Sherman–Morrison identity yields the fundamental deletion formula

$$\beta - \beta_{(i)} = \frac{(X^TX)^{-1}x_ie_i}{1 - h_{ii}}, \tag{9}$$

from which all the classical influence statistics follow without refitting. Cook's distance (Cook, 1977) is the squared displacement of the coefficient vector in the metric of its estimated covariance,

$$D_i = \frac{\left(\beta - \beta_{(i)}\right)^T X^T X\left(\beta - \beta_{(i)}\right)}{ps^2} = \frac{r_i^2}{p} \cdot \frac{h_{ii}}{1 - h_{ii}}. \tag{10}$$

DFFITS (Belsley, Kuh and Welsch, 1980) is the scaled change in the i-th fitted value,

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{s_{(i)}\sqrt{h_{ii}}} = t_i\sqrt{\frac{h_{ii}}{1 - h_{ii}}}, \tag{11}$$

10.48047/jocaaa.2026.35.06.31

with the size-adjusted cut-off $2\sqrt{p/n}$. DFBETAS is the scaled change in the j-th coefficient, with c_{jj} the j-th diagonal element of $(X^TX)^{-1}$,

$$DFBETAS_{ij} = \frac{\beta_j - \beta_{j(i)}}{s_{(i)}\sqrt{c_{jj}}}. \tag{12}$$

with the size-adjusted cut-off $2/\sqrt{n}$.

The covariance ratio compares the generalised variance of the coefficient estimates with and without the case,

$$COVRATIO_i = \frac{det\left[s_{(i)}^2\left(X_{(i)}^T X_{(i)}\right)^{-1}\right]}{det[s^2(X^TX)^{-1}]} = \frac{1}{\left[\frac{n-p-1+t_i^2}{n-p}\right]^p (1-h_{ii})}, \tag{13}$$

and values outside $1 \pm 3p/n$ are taken to indicate a material effect on precision. A value above one means that the case improves precision (typically a good leverage point); a value below one means that it degrades precision (typically a vertical outlier).

3.5 The proposed six-stage identification procedure

The quantities above are combined in the following staged procedure. The stages are sequential in the sense that each uses the output of the preceding one, but the final classification in Stage 5 is made jointly, not by successive filtering.

1. **Stage 1 – Preliminary model fitting.** Fit the OLS model, examine residual plots against fitted values and each predictor, and check for non-linearity, heteroscedasticity and non-normality. Diagnostics for atypical cases are only meaningful once gross model misspecification has been excluded, because a curved relationship fitted by a straight line will generate apparent outliers at the extremes of the predictor range.
2. **Stage 2 – Residual diagnostics.** Compute the externally studentized residuals and flag cases whose absolute value exceeds the Bonferroni critical value. Record the PRESS residuals for later comparison with the ordinary residuals; a large ratio of PRESS residual to ordinary residual indicates that the case is being accommodated by the fit.
3. **Stage 3 – Leverage diagnostics.** Compute the leverage values and the robust distances of the predictor rows from a high-breakdown location and scatter estimate (the minimum covariance determinant estimator of Rousseeuw, 1984). Flag cases for which the classical leverage exceeds twice its mean or the robust distance exceeds the square root of the upper 97.5 percent point of the chi-squared distribution on the number of non-constant predictors. The robust distance is needed because leverage computed from

10.48047/jocaaa.2026.35.06.31

the classical covariance matrix is itself subject to masking when several remote predictor rows occur together.

4. **Stage 4 – Influence diagnostics.** Compute Cook's distance, DFFITS, DFBETAS and the covariance ratio for every case, together with the residuals from a high-efficiency MM-estimator (Yohai, 1987). Flag cases according to the cut-offs given above, and separately identify the coefficients most affected through DFBETAS.
5. **Stage 5 – Joint assessment.** Classify every case into one of the four regions of the residual–leverage plane (regular, vertical outlier, good leverage, bad leverage) using the robust residuals and robust distances, and superimpose the deletion-influence evidence. The classification rule and its justification are developed in Section 4.4. Where the flagged cases form a group, replace the single-case deletion statistics by their group-deletion counterparts as described in Section 4.3.
6. **Stage 6 – Sensitivity analysis.** Refit the model with the identified cases removed and, separately, with the MM-estimator, and compare the full and reduced coefficient vectors, their standard errors, the residual standard deviation and the coefficient of determination. The comparison statistic of Section 4.5 quantifies the total change. Decisions about retention or removal are then taken according to the protocol of Section 4.6.

4. Main Research: Joint Structure of Residual, Leverage and Influence

4.1 Decomposition of the influence measures

Equation (10) shows that Cook's distance is the product of two factors: a residual factor r_i^2/p that depends on the response, and a leverage factor $h_{ii}/(1 - h_{ii})$ that depends only on the predictors. Exactly the same factorisation holds for DFFITS, whose square is

$$DFFITS_i^2 = t_i^2 \cdot \frac{h_{ii}}{1 - h_{ii}} \approx pD_i, \tag{14}$$

the approximation being exact when $s_{(i)}$ is replaced by s. It follows immediately that a large value of D_i can arise in three qualitatively different ways: (i) a large studentized residual at a case of moderate leverage; (ii) a moderate residual at a case of very high leverage; (iii) a large residual at a case of high leverage. Cases of type (i) are vertical outliers, cases of type (ii) are leverage points that are only mildly discordant, and cases of type (iii) are bad leverage points. A single number cannot distinguish them, yet the appropriate remedial action is different in each case. The leverage factor $h_{ii}/(1 - h_{ii})$ is unbounded as h_{ii} approaches one, which

10.48047/jocaaa.2026.35.06.31

explains why moderate residuals at extreme leverage can produce the largest Cook's distances in a data set.

The reverse implication is more important and less often stated. Rearranging (10),

$$r_i^2 = pD_i \frac{1 - h_{ii}}{h_{ii}}, \tag{15}$$

so that for fixed influence D_i the studentized residual shrinks to zero as the leverage increases. A case can therefore be highly influential and have a residual that passes every outlier test. Conversely, from (8) and (9), the PRESS residual $e_{(i)}$ is inflated relative to e_i by the factor $1/(1 - h_{ii})$, so that the discrepancy between a case and the model fitted without it becomes visible only when the case is removed. This is the algebraic content of the statement that high-leverage points 'pull the fit towards themselves'.

4.2 Why no single measure is adequate

The four classical quantities examine four different projections of the same deletion vector, and each is blind in a definite direction. We list the failure modes explicitly.

- **Studentized residuals** detect vertical outliers at low or moderate leverage, but by (15) they fail for discordant cases at high leverage. They are also computed from the OLS fit, which may already have been rotated towards a group of bad leverage points; in that situation the residuals of the bad points are small and the residuals of some regular points are large (swamping).
- **Leverage** identifies remoteness in predictor space but says nothing about the response. A good leverage point with a tiny residual has the same leverage as a bad leverage point with a large one. Moreover, the classical leverage is computed from the sample covariance of the predictors, which is itself inflated by a cluster of remote rows; a cluster of several leverage points can therefore have individually unremarkable leverage values (masking in the predictor space).
- **Cook's distance and DFFITS** combine residual and leverage but, as shown in Section 4.1, cannot indicate which factor is responsible. They are also single-case statistics: when two cases are jointly influential but each alone is not, both deletion vectors are small. A well-known configuration is a pair of bad leverage points located close together in predictor space; deleting either one leaves the other to hold the fitted line in place.
- **DFBETAS** are informative about which coefficients move, but a case that changes many coefficients slightly may have no individual DFBETAS above the cut-off while

10.48047/jocaaa.2026.35.06.31

having a large Cook's distance, because Cook's distance aggregates the changes in the metric of the coefficient covariance.

- **The covariance ratio** is dominated by the leverage term for good leverage points and by the residual term for vertical outliers, and the two effects can cancel: a bad leverage point with a large residual and high leverage can have a covariance ratio close to one.

The conclusion is that a case must be located simultaneously in the residual dimension, the leverage dimension and the influence dimension. These three dimensions are not independent—equation (10) is precisely the constraint between them—but the constraint holds only for the case-deleted, single-observation quantities computed from the OLS fit. When the OLS fit is itself distorted, or when cases occur in groups, the constraint is broken and the three sources of evidence can disagree. It is in this disagreement that diagnostic information lies.

4.3 Masking, swamping and group deletion

Let A be a subset of m cases with predictor rows X_A and residuals e_A, and let H_{AA} be the corresponding $m \times m$ block of the hat matrix. The group-deletion generalisation of (9) is

$$\hat{\beta} - \hat{\beta}_{(A)} = (X^T X)^{-1} X_A^T (I - H_{AA})^{-1} e_A, \tag{16}$$

and the corresponding group Cook statistic is

$$D_A = \frac{e_A^T (I - H_{AA})^{-1} H_{AA} (I - H_{AA})^{-1} e_A}{ps^2}. \tag{17}$$

When $m = 1$ these reduce to (9) and (10). The essential difference is the presence of the off-diagonal elements of H_{AA}. If two cases i and k have predictor rows close together, then h_{ik} is close to $min(h_{ii}, h_{kk})$, the matrix $I - H_{AA}$ is nearly singular, and D_A can be very large even when D_i and D_k are both small. This is the mechanism of masking: the single-case statistic for i is computed with case k still in the fit, and k holds the fit in the position that i would otherwise have pulled it to.

Swamping is the complementary phenomenon. Because the OLS residuals satisfy $X^T e = 0$, a group of bad leverage points that rotates the fitted plane forces the residuals of regular cases in the opposite direction to become large. These regular cases then exhibit large studentized residuals, large Cook's distances and DFFITS values beyond the cut-offs, and will be flagged by every single-case statistic. Removing them makes the fit worse, not better, because the true source of the distortion remains.

Two consequences follow for the methodology. First, leverage and residual screening in Stages 2 and 3 must use quantities that are not themselves derived from the contaminated OLS fit; this

10.48047/jocaaa.2026.35.06.31

is why robust distances and robust residuals are used alongside the classical values. Second, whenever the classical screens flag more than one case, the joint influence of the flagged set must be assessed with (17), and the assessment should be repeated for subsets of the flagged set. Exhaustive subset search is infeasible for large m; the forward-search ordering of Atkinson and Riani (2000), or the eigen-analysis of the influence matrix of Peña and Yohai (1995), provides the candidate subsets in practice.

4.4 Joint classification rule

Let $\tilde{t}_i$ denote the residual of case i from the MM-fit divided by the robust scale estimate, and let RD_i denote the robust distance of x_i (excluding the intercept column) from the MCD location in the MCD metric. Under the model with Gaussian errors and approximately elliptical predictors, $\tilde{t}_i$ is approximately standard normal and RD_i^2 is approximately chi-squared on $q = p - 1$ degrees of freedom for the regular cases. Define the thresholds

$$c_r = 2.5, \quad c_L = \sqrt{\chi_q^2(0.975)}, \tag{18}$$

following Rousseeuw and van Zomeren (1990). Each case is assigned to one of four classes:

Class	Robust residual	Robust distance	Interpretation	Typical deletion evidence
R (regular)	$\lvert\tilde{t}\rvert \le c_r$	$RD \le c_L$	Consistent with model, central in predictor space	Small D, small |DFFITS|, COVRATIO ≈ 1
V (vertical outlier)	$\lvert\tilde{t}\rvert > c_r$	$RD \le c_L$	Discordant response at ordinary leverage	Large |t|, moderate D, COVRATIO < 1
G (good leverage)	$\lvert\tilde{t}\rvert \le c_r$	$RD > c_L$	Remote in predictor space but consistent with model	Small |t|, D may be moderate, COVRATIO > 1
B (bad leverage)	$\lvert\tilde{t}\rvert > c_r$	$RD > c_L$	Remote and discordant	Small OLS |t| possible (masked), large D_A on group deletion

Table 1. Four-way classification of observations in the robust residual–distance plane and the corresponding deletion evidence.

The classification is then reconciled with the deletion statistics of Stage 4. Three patterns of disagreement are diagnostic. If a case is in class B but its OLS Cook's distance is small, masking is present and the group statistic (16) must be computed for the set of all B cases. If a case is in class R but has a large OLS studentized residual or Cook's distance, it is being swamped by the B cases, and it should not be removed. If a case is in class G and has a large Cook's distance,

10.48047/jocaaa.2026.35.06.31

the influence is entirely due to leverage: its removal will widen the confidence intervals of the coefficients to which it contributes without changing their centre, and it should ordinarily be retained.

For reporting purposes the three sources of evidence may be summarised in a single index that records, for each case, how far it lies beyond the thresholds in each dimension. Writing a_+ for the positive part of a, define

$$J_i = \left[\frac{|t_i|}{c_r} - 1\right]_+^2 + \left[\frac{RD_i}{c_L} - 1\right]_+^2 + \left[\frac{D_i}{c_D} - 1\right]_+^2, \quad (19)$$

with c_D the chosen Cook cut-off (the 50 percent point of $F_{p,n-p}$ or $4/n$). The index is zero for cases that are unremarkable in all three dimensions and grows quadratically with excess in any one. It is intended as a ranking device for the analyst's attention, not as a test statistic: its three components are not independent, its sampling distribution is not tractable, and it deliberately loses the information about which dimension is responsible, which must be recovered from Table 1. We do not claim novelty for the index—it is a straightforward combination of quantities already in the literature—and its role in the procedure is purely organisational.

4.5 Sensitivity comparison of full and reduced fits

Let $\hat{\beta}_F$ and $\hat{\beta}_R$ be the OLS estimates from the full data and from the data with the set A of identified cases removed. The total change is summarised by the generalised Cook statistic

$$\Delta_A = \frac{(\hat{\beta}_F - \hat{\beta}_R)^T X^T X (\hat{\beta}_F - \hat{\beta}_R)}{p s_R^2}, \quad (20)$$

which is (17) evaluated with the reduced-fit variance estimate, and by the componentwise standardised changes

$$\delta_j = \frac{\hat{\beta}_{Fj} - \hat{\beta}_{Rj}}{se\,\hat{\beta}_{Rj}}, \quad j = 1, \ldots, p. \quad (21)$$

A value of Δ_A exceeding the 50 percent point of $F_{p,n-m-p}$ indicates that the removed cases move the coefficient vector to the edge of a 50 percent confidence region of the reduced fit (Cook, 1977); a value of $|\delta_j|$ exceeding one indicates that the j-th coefficient moves by more than its standard error. Alongside these, the ratio s_R/s_F and the change in the coefficient of determination indicate the effect on model fit, and the ratio of the standard errors indicates the effect on precision. The comparison should also be made between the full OLS fit and the MM-fit; if the MM-fit and the reduced OLS fit agree closely, the identification in Stage 5 has

10.48047/jocaaa.2026.35.06.31

captured the cases responsible for the discrepancy, and if they do not, further cases remain to be found.

The effect of contamination on inference can be read from these quantities. A vertical outlier inflates the residual variance and hence all standard errors, widening confidence intervals and reducing the power of tests, while leaving the coefficient estimates approximately unbiased if its leverage is small. A bad leverage point biases the coefficients in the direction of the point and may simultaneously reduce the standard errors, so that a biased coefficient appears precisely estimated. A good leverage point reduces the standard errors of the coefficients to which it contributes without biasing them. These three effects on inference correspond exactly to the three non-regular classes of Table 1.

4.6 Decision protocol

The final stage of the framework is a decision about each non-regular case. Because an influential observation is a statement about the sensitivity of the fit and not about the correctness of the data, the decision cannot be made from the diagnostics alone. The following protocol is proposed.

1. **Verify.** Every case in classes V and B should first be checked against its source. Transcription errors, unit errors and coding errors are common and, when confirmed, the correct value should be substituted or the case removed with a documented reason.
2. **Retain good leverage points.** A case in class G is consistent with the model and improves precision. It should be retained unless there is subject-matter reason to believe that the model does not extend to that region of the predictor space, in which case the issue is one of model scope rather than of a bad observation.
3. **Investigate bad leverage points as evidence about the model.** A case in class B that is verified as correct is evidence that the linear model fails in a region of the predictor space. The appropriate response is often a change of model—a transformation, an interaction, a non-linear term—rather than removal. Removal is justified only when the case lies outside the population of interest.
4. **Down-weight rather than delete when the source is unknown.** When a case in class V or B cannot be verified and there is no subject-matter reason to exclude it, reporting the MM-estimate alongside the OLS estimate is preferable to deletion. The robust estimate down-weights the case continuously according to its residual, avoids the discreteness of an all-or-nothing decision, and does not require the analyst to assert that the case is wrong.

10.48047/jocaaa.2026.35.06.31

5. **Report both fits.** Whatever decision is taken, the full-data fit, the reduced or robust fit and the sensitivity comparison of Section 4.5 should all be reported, so that the reader can judge how much of the conclusion depends on the treatment of a small number of cases.

The protocol makes explicit that the diagnostic framework of Sections 4.1–4.5 answers the question of which cases matter and why, and that the question of what to do about them is answered by combining that information with knowledge of the data-generating process. The most serious errors in practice arise from removing cases because a statistic exceeded a cut-off, and from retaining cases because no statistic did.

4.7 Summary of the framework

Method	Vertical outlier	Leverage	Influence	Resists masking	Resists swamping	Cost
Studentized residual	Yes (low leverage)	No	No	No	No	One OLS fit
Leverage h_{ii}	No	Yes (isolated)	No	No	n/a	One OLS fit
Cook's distance / DFFITS	Partly	Partly	Yes (single case)	No	No	One OLS fit
DFBETAS	Partly	Partly	Yes (per coefficient)	No	No	One OLS fit
Covariance ratio	Partly	Partly	Precision only	No	No	One OLS fit
Group deletion D_A	Yes	Yes	Yes (subset)	Yes (given subset)	Yes (given subset)	Combinatorial
Robust residual–distance plot	Yes	Yes	Indirect	Yes	Yes	MM + MCD fits
Proposed joint procedure	Yes	Yes	Yes	Yes	Yes	OLS + MM + MCD + subset D_A

Table 2. Qualitative comparison of diagnostic approaches within the proposed framework.

5. Conclusion

This paper has set out, in a single notation, the algebraic structure that links the residual, the leverage and the deletion influence of an observation in the linear regression model, and has

10.48047/jocaaa.2026.35.06.31

used that structure to explain why each of the classical diagnostic measures fails in a definite and predictable way. Cook's distance and DFFITS factorise into a residual component and a leverage component, so a large value is ambiguous; the studentized residual of a high-leverage case is shrunk towards zero for any fixed level of influence, so discordant leverage points pass outlier tests; the single-case deletion formulae depend on the remaining cases holding the fit in place, so groups of atypical cases mask one another; and the orthogonality of OLS residuals to the predictors forces regular cases to absorb the distortion caused by bad leverage points, so that innocent cases are swamped.

The proposed six-stage procedure responds to these findings by screening in each dimension with quantities that are not themselves derived from the contaminated fit, by classifying every case jointly in the robust residual–distance plane, by reconciling the classification with the deletion statistics so that masking and swamping are recognised as patterns of disagreement, and by quantifying the sensitivity of the coefficients, their precision and the fit to the treatment of the identified cases. A decision protocol translates the classification into action, with the central principle that an influential observation is a statement about the fit and not a verdict on the data.

The theoretical contribution is the systematic derivation of the failure modes and the group-deletion mechanism of masking; the methodological contribution is the integration of classical deletion diagnostics with robust classification into one staged procedure with an explicit decision rule; the applied contribution is a set of concrete recommendations for practitioners. The framework is confined to the linear model with independent homoscedastic errors and to settings in which the number of predictors is small relative to the number of observations. Extension to generalised linear and mixed models, where the hat matrix depends on the response through the working weights, to time-series regression, where deletion breaks the dependence structure, and to high-dimensional and penalised regression, where the leverage values are no longer well-defined in the classical sense, is left for subsequent work, as is the empirical assessment of the procedure by simulation and on real data.

References

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

Anscombe, F. J. and Tukey, J. W. (1963). The examination and analysis of residuals. Technometrics, 5(2), 141–160.

10.48047/jocaaa.2026.35.06.31

Atkinson, A. C. (1985). Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis. Oxford: Clarendon Press.

Atkinson, A. C. (1986). Masking unmasked. Biometrika, 73(3), 533–541.

Atkinson, A. C. and Riani, M. (2000). Robust Diagnostic Regression Analysis. New York: Springer.

Barnett, V. and Lewis, T. (1994). Outliers in Statistical Data, 3rd edition. Chichester: Wiley.

Beckman, R. J. and Cook, R. D. (1983). Outlier.....s. Technometrics, 25(2), 119–149.

Behnken, D. W. and Draper, N. R. (1972). Residuals and their variance patterns. Technometrics, 14(1), 101–111.

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

Chatterjee, S. and Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. Statistical Science, 1(3), 379–393.

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

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

Cook, R. D. (1979). Influential observations in linear regression. Journal of the American Statistical Association, 74(365), 169–174.

Cook, R. D. (1986). Assessment of local influence (with discussion). Journal of the Royal Statistical Society, Series B, 48(2), 133–169.

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

Davies, L. and Gather, U. (1993). The identification of multiple outliers (with discussion). Journal of the American Statistical Association, 88(423), 782–792.

Draper, N. R. and John, J. A. (1981). Influential observations and outliers in regression. Technometrics, 23(1), 21–26.

Draper, N. R. and Smith, H. (1966). Applied Regression Analysis. New York: Wiley.

Draper, N. R. and Smith, H. (1998). Applied Regression Analysis, 3rd edition. New York: Wiley.

Ellenberg, J. H. (1976). Testing for a single outlier from a general linear regression. Biometrics, 32(3), 637–645.

Fox, J. (1991). Regression Diagnostics. Newbury Park, CA: Sage.

10.48047/jocaaa.2026.35.06.31

Habshah, M., Norazan, M. R. and 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(5), 507–520.

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

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. Journal of the American Statistical Association, 88(424), 1264–1272.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69(346), 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). Robust Statistics: The Approach Based on Influence Functions. New York: Wiley.

Hawkins, D. M. (1980). Identification of Outliers. London: Chapman and Hall.

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. The American Statistician, 32(1), 17–22.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. Annals of Statistics, 1(5), 799–821.

Huber, P. J. (1981). Robust Statistics. New York: Wiley.

Hubert, M., Rousseeuw, P. J. and Van Aelst, S. (2008). High-breakdown robust multivariate methods. Statistical Science, 23(1), 92–119.

Imon, A. H. M. R. (2005). Identifying multiple influential observations in linear regression. Journal of Applied Statistics, 32(9), 929–946.

Kianifard, F. and Swallow, W. H. (1996). A review of the development and application of recursive residuals in linear models. Journal of the American Statistical Association, 91(433), 391–400.

Kutner, M. H., Nachtsheim, C. J. and Neter, J. (2004). Applied Linear Regression Models, 4th edition. New York: McGraw-Hill/Irwin.

Lawrance, A. J. (1988). Regression transformation diagnostics using local influence. Journal of the American Statistical Association, 83(404), 1067–1072.

Lawrance, A. J. (1995). Deletion influence and masking in regression. Journal of the Royal Statistical Society, Series B, 57(1), 181–189.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). Robust Statistics: Theory and Methods. Chichester: Wiley.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2012). Introduction to Linear Regression Analysis, 5th edition. Hoboken, NJ: Wiley.

10.48047/jocaaa.2026.35.06.31

Myers, R. H. (1990). Classical and Modern Regression with Applications, 2nd edition. Boston: PWS-Kent.

Peña, D. (2005). A new statistic for influence in linear regression. Technometrics, 47(1), 1–12.

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(1), 145–156.

Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American Statistical Association, 79(388), 871–880.

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

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85(411), 633–639.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle and R. D. Martin (eds), Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics 26, pp. 256–272. New York: Springer.

Tietjen, G. L., Moore, R. H. and Beckman, R. J. (1973). Testing for a single outlier in simple linear regression. Technometrics, 15(4), 717–721.

Velleman, P. F. and Welsch, R. E. (1981). Efficient computing of regression diagnostics. The American Statistician, 35(4), 234–242.

Weisberg, S. (1985). Applied Linear Regression, 2nd edition. New York: Wiley.

Wu, X. and Luo, Z. (1993). Second-order approach to local influence. Journal of the Royal Statistical Society, Series B, 55(4), 929–936.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. Annals of Statistics, 15(2), 642–656.