28. Measures of Influence and Sensitivity in Linear Regression

This chapter reviews diagnostic procedures for detecting outliers and influential observations in linear regression. First, the statistics for detecting single outliers and influential observations are presented, and their limitations for multiple outliers in high-leverage situations are discussed; second, diagnostic procedures designed to avoid masking are shown. We comment on the procedures by *Hadi* and *Smirnoff* [28.1,2], *Atkinson* [28.3] and *Swallow* and *Kianifard* [28.4] based on finding a clean subset for estimating the parameters and then increasing its size by incorporating new homogeneous observations one by one, until a heterogeneous observation is found. We also discuss procedures for detecting high-leverage outliers in large data sets based on eigenvalue analysis of the influence and sensitivity matrix, as proposed by *Peña* and *Yohai* [28.5, 6]. Finally we show that the joint use of simple univariate statistics, as predictive residuals, and Cook's distances, jointly with the sensitivity statistic proposed by *Peña* [28.7] can be a useful diagnostic tool for large high-dimensional data sets.

Data often contain outliers or atypical observations. Outliers are observations which are heterogeneous with the rest of the data, due to large measurement errors, different experimental conditions or unexpected effects. Detecting these observations is important because they can lead to new discoveries. For instance, penicillin was found because Pasteur, instead of ignoring an outlier, tried to understand the reason for this atypical effect. As *Box* [28.8] has emphasized "every operating system supplies information on how it can be improved and if we use this information it can be a source of continuous improvement". A way in which this information appears is by outlying observations, but in many engineering processes these observations are not easy to detect. For instance, in a production process a large value in one of the variables we monitor may be due, among other causes, to: (1) a large value of one of the input control variables; (2) an unexpected interaction among the input variables; (3) a large measurement error due to some defect in the measurement instrument. In the first case, the outlying observations may provide no new information about the performance of the process but in the second case may lead to a potentially useful discovery and in the third, to an improvement of the process control. A related problem is to avoid the situation where these outliers affect the estimation of the statistical model and this is the aim of robust estimation methods.

This chapter discusses outliers, influential and sensitive observations in regression models and presents methods to detect them. Influential observations are those which have a strong influence on the global properties of the model. They are obtained by modifying the weights attached to each case, and looking at the standardized change of the parameter vector or the vector of forecasts. Influence is a global analysis. Sensitive observations can be declared outliers or not by small modifications in the sample. Sensitivity is more a local concept. We delete each sample point in turn and look at the change that these modifications produce in the forecast of a single point. We will see that influence

and sensitivity are important concepts for understanding the effect of data in building a regression model and in finding groups of outliers.

Many procedures are available to identify a single outlier or an isolated influential point in linear regression. The books of *Belsley* et al. [28.9], *Hawkins* [28.10], *Cook* and *Weisberg* [28.11], *Atkinson* [28.12], *Chatterjee* and *Hadi* [28.13] *Barnett* and *Lewis* [28.14] and *Atkinson* and *Riani* [28.15] present good analyses of this problem. To identify outliers and to measure influence the point can be deleted, as proposed by *Cook* [28.16] and *Belsley* et al. [28.9], or its weight decreased, as in the local influence analysis introduced by *Cook* [28.17]. See *Brown* and *Lawrence* [28.18] and *Suárez Rancel* and *González Sierra* [28.19] for a review of local influence in regression and many references, and *Hartless* et al. [28.20] for recent results on this approach. A related way to analyze influence has been proposed by *Critchley* et al. [28.21] by an extension of the influence-curve methodology. The detection of influential subsets or multiple outliers is more difficult, due to the masking and swamping problems. Masking occurs when one outlier is not detected because of the presence of others; swamping happens when a non-outlier is wrongly identified due to the effect of some hidden outliers, see *Lawrance* [28.22]. Several procedures have been proposed for dealing with multiple outliers, see *Hawkins*, *Bradu* and *Kass* [28.23], *Gray* and *Ling* [28.24], *Marasinghe* [28.25], *Kianifard* and *Swallow* [28.26, 27], *Hadi* and *Simonoff* [28.1, 2], *Atkinson* [28.3, 28] and *Swallow* and *Kianifard* [28.4]. A different analysis for detecting groups of outliers by looking at the eigenvectors of an influence matrix was presented by *Peña* and *Yohai* [28.5]. These authors later proposed [28.6] the sensitivity matrix as a better way to find interesting groups of data, and from this approach *Peña* [28.7] has proposed a powerful diagnostic statistic for detecting groups of outliers.

We do not discuss in this chapter, due to lack of space, robust regression methods and only refer to them when they are used as a first step in a diagnosis procedure. See *Huber* [28.29] for a good discussion of the complementary role of diagnosis and robustness. For robust estimation in regression see *Rousseeuw* and *Leroy* [28.30] and *Maronna*, *Martin* and *Yohai* [28.31]. Robust estimation of regression models has also received attention in the Bayesian literature since the seminal article of *Box* and *Tiao* [28.32]. See *Justel* and *Peña* [28.33] for a Bayesian approach to this problem and references.

The paper is organized as follows. In Sect. 28.1 we present the model and the notation, and define the main measures which will be used for outlier and influence analysis. In Sect. 28.2 we review procedures for detecting single outliers and influential observations in regression. In Sect. 28.3 we discuss the multiple-outlier problem and two types of diagnostic procedures, those based on an initial clean subset and those based on eigenvalue analysis of some diagnostic matrices. In Sect. 28.4 we introduce a simple statistic which can be used for diagnostic analysis of a large data set, avoiding the masking problem. Section 28.5 includes an example of the use of diagnostic methods for detecting groups of outliers and Sect. 28.6 contains some concluding remarks.

28.1 The Leverage and Residuals in the Regression Model

We assume that we have observed a sample of size n of a random variable $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and a set of $p-1$ explanatory variables which are linearly related by

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \tag{28.1}$$

where the u_i are the measurement errors, which will be independent normal zero-mean random variables with variance σ^2, and $\boldsymbol{u} = (u_1, \ldots, u_n)'$. The $\boldsymbol{x}_i = (1, x_{2i}, \ldots, x_{pi})$ are numerical vectors in R^p and we will denote by $\mathbf{X}$ the $n \times p$ matrix of rank p whose i-th row is $\boldsymbol{x}_i'$. Then, the least-squares estimate of $\boldsymbol{\beta}$ is obtained by projecting the vector $\boldsymbol{y}$ onto the space generated by the columns of $\mathbf{X}$, which leads to

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$$

and the vector of fitted values, $\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_n)'$, is given by

$$\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\boldsymbol{y}, \tag{28.2}$$

where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

is the idempotent projection matrix. The vector orthogonal to the space generated by the $\mathbf{X}$ variables is the residual vector, $\boldsymbol{e} = (e_1, \ldots, e_n)'$, which is defined

by

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\mathbf{I} - \mathbf{H})\boldsymbol{y} \tag{28.3}$$

and we will let $\widehat{s}^2_{\mathrm{R}} = \boldsymbol{e}'\boldsymbol{e}/(n-p)$ be the estimated residual variance.

From (28.3), inserting $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{u}$ instead of $\boldsymbol{y}$ and using $\mathbf{HX} = \mathbf{X}$, we obtain the relationship between the residuals and the measurement errors, $\boldsymbol{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{u}$. Thus, each residual is a linear combination of the measurement errors. Letting $h_{ij} = \boldsymbol{x}_i' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \boldsymbol{x}_j$ be the elements of the matrix, $\mathbf{H}$, we have

$$e_i = u_i - \sum_{j=1}^{n} h_{ij} u_j \tag{28.4}$$

and, if the second term is small, the residual e_i will be close to the measurement error, u_i. The variance of this second term is

$$Var(\sum_{j=1}^{n} h_{ij} u_j) = \sigma^2 \sum_{j=1}^{n} h_{ij}^2 = \sigma^2 h_{ii}$$

and if h_{ii}, the diagonal term of $\mathbf{H}$, is large, the difference between the residual and the measurement error can be large. The values h_{ii} are called the leverage of the observation and measure the discrepancy of each observation $\boldsymbol{x}_i$ with respect to the mean of the explanatory variables. It can be shown (see for instance [28.11] p. 12) that

$$h_{ii} = \boldsymbol{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_i = \frac{1}{n}\left[1 + \left(\widetilde{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}\right)' \mathbf{S}_{xx}^{-1} (\widetilde{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})\right],$$

where $\widetilde{x}_h = (x_{2h}, \ldots, x_{ph})$ does not include the constant term, $\bar{\boldsymbol{x}}$ is the vector of means of the $p-1$ explanatory variables and $\mathbf{S}_{xx}$ is their covariance matrix. Note that, if the variables were uncorrelated, h_{ii} would be the sum of the standardized distances $[(x_{ij} - x_j)/s_j]^2$. As $\sum_{i=1}^{n} h_{ii} = \mathrm{tr}(\mathbf{H}) = p$, the average value of the leverage is $\bar{h} = \sum h_{ii}/n = p/n$, and it can be shown that $1/n \le h_{ii} \le 1$. From (28.4) we conclude that the residual will be close to the measurement error for those observations close to the center of the explanatory data, where $h_{ii} \simeq 1/n$, but will be very different for the extreme points where $h_{ii} \simeq 1$. Note that the residual covariance matrix is

$$\begin{aligned} Var(\boldsymbol{e}) &= E[\boldsymbol{e}\boldsymbol{e}'] = E[(\mathbf{I} - \mathbf{H})\boldsymbol{u}\boldsymbol{u}'(\mathbf{I} - \mathbf{H})] \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) \end{aligned} \tag{28.5}$$

and $Var(e_i) = \sigma^2(1 - h_{ii})$, which will be large when $h_{ii} \simeq 1/n$, and close to zero if $h_{ii} \simeq 1$. As the mean of the residuals is zero if the variance of e_i is very small this implies that its value will be close to zero, whatever the value of u_i.

The problem that each residual has a different variance leads to the definition of the standardized residuals, given by

$$r_i = \frac{e_i}{\widehat{s}_{\mathrm{R}}\sqrt{1 - h_{ii}}} \tag{28.6}$$

which will have variance equal to one. A third type of useful residuals are the predictive, deleted, or out-of-sample residuals, defined by $e_{(i)} = y_i - \widehat{y}_{i(i)}$, where $\widehat{y}_{i(i)}$ is computed in a sample with the i-th observation deleted. It can be shown that

$$e_{(i)} = \frac{e_i}{(1 - h_{ii})} \tag{28.7}$$

and the variance of these predictive residuals is $\sigma^2/(1 - h_{ii})$. If we estimate σ^2 by $\widehat{s}^2_{\mathrm{R}(i)}$, the residual variance in a regression which does not include the i-th observation, the standardization of the predictive residual leads to the Studentized residual, defined by

$$\widehat{t}_i = \frac{e_i}{\widehat{s}_{\mathrm{R}(i)}\sqrt{1 - h_{ii}}} \tag{28.8}$$

which has a Student t distribution with $n - p - 1$ degrees of freedom. An alternative useful expression of these residuals is based on $h_{ii(i)} = \boldsymbol{x}_i'(\mathbf{X}_{(i)}\mathbf{X}_{(i)})^{-1}\boldsymbol{x}_i = h_{ii}/(1 - h_{ii})$, where $\mathbf{X}_{(i)}$ is the $(n-1) \times p$ matrix without the row $\boldsymbol{x}_i'$, and therefore, we have the alternative expression:

$$\widehat{t}_i = \frac{e_{(i)}}{\widehat{s}_{\mathrm{R}(i)}\sqrt{1 + h_{ii(i)}}}. \tag{28.9}$$

28.2 Diagnosis for a Single Outlier

28.2.1 Outliers

If one observation, y_h, does not follow the regression model, either because its expected value is not $\boldsymbol{x}_h'\boldsymbol{\beta}$, or its conditional variance is not σ^2, we will say that it is an outlier. These discrepancies are usually translated to the residuals. For instance, if the observation has been generated by a different model, $g(\boldsymbol{x}_h') + u_h$, then

$$e_h = g(\boldsymbol{x}_h') - \boldsymbol{x}_h'\hat{\boldsymbol{\beta}} + u_h$$

and the deviation $\left|g(\boldsymbol{x}_h')-\boldsymbol{x}_h'\hat{\boldsymbol{\beta}}\right|$ will be larger than $\left|\boldsymbol{x}_h'(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right|$. However, we may not detect this observation because of the key role of the variable $\boldsymbol{x}_h'$. Suppose, in order to simplify, that we write $g(\boldsymbol{x}_h')=\boldsymbol{x}_h'\boldsymbol{\alpha}$, that is, the data is also generated by a linear model but with different parameter values. Then, even if $\boldsymbol{\alpha}$ is very different from $\boldsymbol{\beta}$, the size of $\left|\boldsymbol{x}_h'(\boldsymbol{\alpha}-\hat{\boldsymbol{\beta}})\right|$ depends on $\boldsymbol{x}_h'$ and the discrepancy between the parameter values would be easier to detect when $\left|\boldsymbol{x}_h'\right|$ is large than when it is small.

When the observation is an outlier because it has a measurement error which comes from a distribution with variance $k\sigma^2$, where $k>1$, we expect that $|u_h|$ will be larger than the rest of the measurement errors. It is intuitive, and it has been formally shown [28.34], that we cannot differentiate between a change in the mean and a change in the variance by using just one observation; also models which assume a change in the variance are equivalent to those which assume shifts in the mean of the observations. Thus, we consider a simple mean-shift model for a single outlier

$$y_h=\boldsymbol{x}_h'\boldsymbol{\beta}+w+u_h,$$

where w is the size of the outlier and u_h is $N(0,\sigma^2)$. A test for outliers can be made by estimating the parameter w in the model

$$y_i=\boldsymbol{x}_i'\boldsymbol{\alpha}+wI_i^{(h)}+u_i,\quad i=1,..,n,$$

where $I_i^{(h)}$ is a dummy variable given by $I_i^{(h)}=1$, when $i=h$ and $I_i^{(h)}=0$, otherwise. We can test for outliers by fitting this model for $h=1,\ldots,n$, and checking if the estimated coefficient $\widehat{w}$ is significant. It is easy to show that:

1. $\widehat{\boldsymbol{\alpha}}=(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}'\boldsymbol{y}_{(i)}=\widehat{\boldsymbol{\beta}}_{(i)}$, the regression parameters are estimated in the usual way, but deleting case $(y_j,\boldsymbol{x}_j)$;
2. $\widehat{w}=y_h-\boldsymbol{x}_h'\widehat{\boldsymbol{\alpha}}$, and therefore the estimated residual at this point, $e_h=y_h-\boldsymbol{x}_h'\widehat{\boldsymbol{\alpha}}-\widehat{w}=0$.
3. The t statistic to check if the parameter $\widehat{w}$ is significant is equal to the Studentized residual, t_h, as defined in (28.8).

Assuming that only one observation is an outlier the test is made by comparing the standardized residual to the maximum of a t distribution with $n-p-2$ degrees of freedom. Often, for moderate n, cases are considered as outliers if their Studentized residuals are larger than 3.5.

28.2.2 Influential Observations

An intuitive way to measure the effect of an observation on the estimated parameters, or on the forecasts, is to delete this observation from the sample and see how this deletion affects the vector of parameters or the vector of forecasts. A measure of the influence of the $i-$th observation on the parameter estimate is given by:

$$D(i)=\frac{(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(i)}')\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(i)})}{p\widehat{s}_{\mathrm{R}}^2},\tag{28.10}$$

which, as the covariance of $\widehat{\boldsymbol{\beta}}$ is $\widehat{s}_{\mathrm{R}}^2(\mathbf{X}'\mathbf{X})^{-1}$, measures the change between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{(i)}$ with relation to the covariance matrix of $\widehat{\boldsymbol{\beta}}$, standardized by the dimension of the vector p. This measure was introduced by *Cook* [28.16]. Of course other standardizations are possible. *Belsley* et al. [28.9] propose using $\widehat{s}_{\mathrm{R}(i)}^2$, the variance of the regression model when the ith observation is deleted, instead of $\widehat{s}_{\mathrm{R}}^2$, and *Diaz-García* and *Gonzalez-Farias* [28.35] have suggested standardizing the vector $(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(i)})$ by its variance, instead of using the variance of $\widehat{\boldsymbol{\beta}}$. See *Cook*, *Peña* and *Weisberg* [28.36] for a comparison of some of these possible standarizations.

Equation (28.10) can also be written as the standardized change in the vector of forecasts:

$$D_i=\frac{\left(\widehat{\boldsymbol{y}}-\widehat{\boldsymbol{y}}_{(i)}\right)'\left(\widehat{\boldsymbol{y}}-\widehat{\boldsymbol{y}}_{(i)}\right)}{p\widehat{s}_{\mathrm{R}}^2},\tag{28.11}$$

where $\widehat{\boldsymbol{y}}_{(i)}=\mathbf{X}\widehat{\beta}_{(i)}=(\hat{y}_{1(i)},\ldots,\hat{y}_{n(i)})'$. Note that from (28.2) we have that $Var(\widehat{y}_i)=\sigma^2h_{ii}$ and as the average value of h_{ii} is p/n, (28.11) is standardized by this average value and by the dimension n of the vector. A third way to measure the influence of the ith point is to compare $\widehat{y}_i$ with $\widehat{y}_{(i)}$, where $\widehat{y}_{(i)}=\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(i)}$. With the usual standardization by the variance we have:

$$D_i=\frac{\left(\widehat{y}_i-\widehat{y}_{(i)}\right)^2}{p\widehat{s}_R^2h_{ii}}\tag{28.12}$$

and, using the relation between the inverse of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}_{(i)}'\mathbf{X}_{(i)}$, we obtain

$$\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{(i)}=(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{x}_i\frac{e_i}{1-h_{ii}}.\tag{28.13}$$

Inserting this into (28.10) it is easy to see that (28.12) is equivalent to (28.10) and (28.11). Also, as from (28.13) we have that

$$\widehat{\boldsymbol{y}}-\widehat{\boldsymbol{y}}_{(i)}=\boldsymbol{h}_i\frac{e_i}{1-h_{ii}},\tag{28.14}$$

where $\boldsymbol{h}_i$ is the $i-$th column of the $\mathbf{H}$ matrix, inserting this expression into (28.11) we obtain a convenient expression for the computation of Cook's statistics:

$$D_i = \frac{r_i^2 h_{ii}}{p(1-h_{ii})}, \tag{28.15}$$

where r_i is the standardized residual given by (28.6). For large n, the expected value of D_i can be approximated by

$$E(D_i) \simeq \frac{h_{ii}}{p(1-h_{ii})}, \tag{28.16}$$

and it will be very different for observations with different leverage.

Cook proposed judging the values of D_i by an $F(p; n-p; 1-\alpha)$, where F is the distribution used in building a confidence region for the $\boldsymbol{\beta}$ parameters. Thus, we may identify points as influential when they are able to move the estimate out of the confidence region for a fixed value of α and declare as influential those observations which verify $D_i \geq F(p; n-p; 1-\alpha)$. This solution is not satisfactory for large sample sizes because it is difficult for any observation to be deemed influential. *Muller* and *Mok* [28.37] have obtained the distribution of the D_i for normal explanatory variables, but this distribution is complicated.

Cook [28.17] proposed a procedure for the assessment of the influence on a vector of parameters θ of minor perturbation of a statistical model. This approach is very flexible and can be used to see the effect of small perturbations which would not normally be detected by deletion of one observation. He suggested introducing a $n \times p$ vector $\boldsymbol{w}$ of case weights and use the likelihood displacement $[L(\hat{\theta}) - L(\hat{\theta}_{\mathrm{w}})]$, where $\hat{\theta}$ is the maximum likelihood (ML) estimator of $\hat{\theta}$, and $\hat{\theta}_{\mathrm{w}}$ is the ML when the case weight $\boldsymbol{w}$ is introduced. Then, he showed that the directions of greatest local change in the likelihood displacement for the linear regression model are given by the eigenvectors linked to the largest eigenvalues of the curvature matrix, $\mathbf{L} = \mathbf{EHE}$, where $\mathbf{E}$ is the vector of residuals. Later, we will see how this approach is related to some procedures for multiple-outlier detection.

28.2.3 The Relationship Between Outliers and Influential Observations

An outlier may or may not be an influential observation and an influential observation may or may not be an outlier. To illustrate this point consider the data in Table 28.1. We will use these data to build four data sets. The first includes cases 1–9 repeated three times, and has sample size $n = 27$. The other three are formed by adding a new observation to this data set. The set (a) is built by adding case 28(a), the set (b) by adding case 28(b) and the set (c) by adding case 28(c). Table 28.2 shows some statistics of these four data sets where (0) refers to the set of 27 observations and (a), (b) and (c) to the sets of 28 observations as defined before. The table gives the values of the estimated parameters, their t statistics in parentheses, the residual standard deviation, the leverage of the added point, the standardized residual for the added point and the value of Cook's statistics.

In set (a) observation 28 is clearly an outlier with a value of the standardized residual of 4.68, but it is not influential, as $D_{28}(\mathrm{a}) = 0.92$, which is a small value. In case (b) the 28-th point is not an outlier, as $r_{28}(\mathrm{b}) = 1.77$ is not significant, but it is very influential, as indicated by the large D_{28} value. Finally, in set (c) the observation is both an outlier, $r_{28} = 4.63$, and very influential, $D_{28} = 13.5$.

Table 28.1 Three sets of data which differ in one observation

Case	1	2	3	4	5	6	7	8	9	(a)	(b)	(c)
x_1	−2	0	2	−4	3	1	−3	−1	4	0	−3	−3
x_2	6.5	7.3	8.3	6.0	8.8	8.0	5.9	6.9	9.5	7.2	9	7.3
y	−1.5	0.5	1.6	−3.9	3.5	0.8	−2.7	−1.3	4.1	5	−1.5	4

Table 28.2 Some statistics for the three regressions fitted to the data in Table 28.1

	$\widehat{\beta}_0$	$t(\widehat{\beta}_0)$	$\widehat{\beta}_2$	$t(\widehat{\beta}_2)$	$\widehat{\beta}_1$	$t(\widehat{\beta}_1)$	$\widehat{s}_{\mathrm{R}}$	h_{28}	r_{28}	D_{28}
(0)	2.38	(0.82)	−0.30	(0.78)	1.12	(6.24)	0.348	–	–	–
(*a*)	13.1	(1.7)	−1.72	(−1.66)	1.77	(3.69)	0.96	0.11	4.68	0.92
(*b*)	−2.74	(−2.9)	0.38	(3.08)	0.80	(13.87)	0.36	0.91	1.77	11.1
(*c*)	−25.4	(−5.41)	3.43	(5.49)	−0.624	(2.22)	0.91	0.65	4.63	13.5

Note that if the leverage is small $h_{ii} \simeq 1/n$, $h_{ii}/(1-h_{ii}) \simeq (n-1)^{-1}$, and by (28.15):

$$D_i = \frac{r_i^2}{p}\left(\frac{1}{n-1}\right),$$

then, if n is large, the observation cannot be influential, whatever the value of r_i^2. On the other hand, high-leverage observations with h_{ii} close to one will have a ratio $h_{ii}/(1-h_{ii})$ that is arbitrarily large and, even if r_i^2 is small, will be influential.

28.3 Diagnosis for Groups of Outliers

The procedures that we have presented in the previous section are designed for a single outlier. We can extend these ideas to multiple outliers as follows. Let I be an index set corresponding to a subset of r data points. The checking of this subset can be done by introducing dummy variables as in the univariate case. Assuming normality, the F test for the hypothesis that the coefficients of the dummy variables are zero is given by

$$F_{r,(n-p-r)} = \frac{\boldsymbol{e}_I'(\mathbf{I}-\mathbf{H}_I)^{-1}\boldsymbol{e}_I}{r\widehat{s}^2_{\mathrm{R}(I)}}$$

where $\boldsymbol{e}_I$ is the vector of least-squares residuals, $\mathbf{H}_I$ the $r\times r$ submatrix of $\mathbf{H}$, corresponding to the set of observations included in I, and $\widehat{s}^2_{\mathrm{R}(I)}$ the residual variance of the regression with the set I deleted. *Cook* and *Weisberg* [28.11] proposed to measure the joint influence of the data points with index in I by deleting the set I and computing, as in the single outlier case,

$$D_I = \frac{(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(I)}')\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(I)})}{p\widehat{s}^2_{\mathrm{R}}},$$

which can also be written as a generalization of (28.15) by $D_I = [\boldsymbol{e}_I'(\mathbf{I}-\mathbf{H}_I)^{-1}\mathbf{H}_I(\mathbf{I}-\mathbf{H}_I)^{-1}\boldsymbol{e}_I]/p\widehat{s}^2_{\mathrm{R}}$. Note that a large value of D_I may be due to a single influential observation included in the set I or a sum of small individual effects of a set of observations that are masking each other. However, in the first case this single observation will be easily identified. Also, a subset of individually highly influential points, whose effect is to cancel each other out, will lead to a small value of D_I; again in this case, the individual effects will be easy to identify. However, to build this measure we should compute all sets of I in the n data, which would be impossible for large I and n.

The procedures for finding multiple outliers in regression can be divided into three main groups. The first is based on robust estimation. If we can compute an estimate that is not affected by the outliers, we can then find the outliers as those cases with large residuals with respect to the robust fit. We present briefly here the least median of squares (LMS) estimate proposed by *Rousseeuw* [28.38], which is used as an initial estimate in some diagnostic procedures based on a clean set, which we will review below. *Rousseeuw* [28.38] proposed generating many possible values of the parameters, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$, finding the residuals associated with each parameter value, $\boldsymbol{e}_i = \boldsymbol{y} - \mathbf{X}\,\boldsymbol{\beta}_i$ $(i = 1, .., N)$, and using the median of these residuals as a robust scale

$$s(\boldsymbol{\beta}_i) = median(e_{1i}^2, \ldots, e_{ni}^2). \qquad (28.17)$$

The value $\boldsymbol{\beta}_i$ that minimizes this robust scale is the LMS estimate. *Rousseeuw* [28.38] generates the parameter values $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$ by resampling, that is, by taking many random samples of size p, $(\mathbf{X}_i, \boldsymbol{y}_i)$, where the matrix $\mathbf{X}_i$ is $p\times p$ and $\boldsymbol{y}_i$ is $p\times 1$, and computing the least-squares estimate (LSE) for each sample, $\boldsymbol{\beta}_i = \mathbf{X}_i^{-1}\boldsymbol{y}_i$. The LMS, although very robust, is not very efficient, and many other robust methods have been proposed to keep high robustness and achieve better efficiency in regression [28.31].

A second class of procedures uses robust ideas to build an initial clean subset and then combine least-squares estimates in clean subsets and diagnosis ideas for outlier detection. Three procedures in this spirit will be presented next; they can be very effective when p and n are not large.

For large data sets with many predictors and high-leverage observations, robust estimates can be very difficult to compute and procedures based on the clean-set idea may not work well, because of the difficulty in selecting the initial subset. The third type of procedures are based on eigenstructure analysis of some diagnostic matrices and are especially useful for large data sets.

28.3.1 Methods Based on an Initial Clean Set

Kianifard and *Swallow* [28.26, 27] proposed to build a *clean* set of observations and check the rest of the data with respect to this set. If the observation closest to the clean set is not an outlier, then the clean set is increased by one observation, and continue to do so until no new

observations can be incorporated into the basic set. The key step in this procedure is to find the initial subset, because if it contains outliers the whole procedure breaks down. These authors proposed using either the predictive or standardized residuals, or a measure of influence such as D_i.

A similar procedure was proposed by *Hadi* and *Simonoff* [28.1, 2]. In [28.2] they recommend building the initial subset using the LMS. The clean set is built by computing this robust estimate and then uses the $h = \left(\frac{n+p+1}{2}\right)$ observations with the smallest residuals with respect to this robust fit to form the initial clean set, which we call M. The procedure continues by fitting a regression model by least squares to this clean set M. Calling $\widehat{\boldsymbol{\beta}}_M$ the estimated LSE parameters and $\widehat{\sigma}_M$ the residual standard deviation, a set of in-sample and out-of-sample residuals is obtained as follows

$$d_i = \frac{\left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_M\right|}{\widehat{\sigma}_M\sqrt{1-\boldsymbol{x}_i'(\mathbf{X}_M'\mathbf{X}_M)^{-1}\boldsymbol{x}_i}}, \quad \text{if } i \in M,$$

$$d_i = \frac{\left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_M\right|}{\widehat{\sigma}_M\sqrt{1+\boldsymbol{x}_i'(\mathbf{X}_M'\mathbf{X}_M)^{-1}\boldsymbol{x}_i}}, \quad \text{if } i \notin M.$$

That is, d_i represents the standardized residual (28.6) for the data in set M and the predictive residual (28.9) for observations outside this set. Then, all of the observations are arranged in increasing order according to d_i. Let s be the size of the set M (which is h in the first iteration, but will change as explained below). If $d_{(s+1)}$ is smaller than some critical value, a new set of size $s+1$ is built with the $s+1$ observations with smallest d values. If $d_{(s+1)}$ is larger than some critical value, all observations out of the set M are declared as outliers and the procedure stops. If $n = s+1$ we stop and declare that there are no outliers in the data. These authors proposed using as critical values those of the t distribution adjusted by Bonferroni, that is $t\left(\frac{\alpha}{2(s+1)}, s-p\right)$.

Atkinson [28.3] proposed a similar approach called the forward search. His idea is again to combine a robust estimate with diagnostic analysis. He computes the LMS estimate but, instead of generating a large set of candidates by random sample, he generates a set of candidate values for $\widehat{\boldsymbol{\beta}}$ by fitting least-squares subsamples of size p, $p+1, \ldots, n$. The procedure is as follows. We start by generating a random sample of size p; let I_p be the indices of the observations selected. Then, we compute the parameters $\widehat{\boldsymbol{\beta}}(p)$ by LSE, and the residual for all cases, $\boldsymbol{e} = \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(p)$. The residuals are corrected by

$$u_i^2 = e_i^2, \qquad i \in I \tag{28.18}$$

$$u_i^2 = e_i^2/(1+h_{ii}), \qquad i \notin I$$

and these residuals u_i^2 are ordered and the smallest $p+1$ are selected. With this new sample of size $m = p+1$ the process is repeated, that is, the parameters are computed by LSE and the residuals to this fit for the n points are obtained. The corrected residuals (28.18) are computed and the process is continued. In this way we obtain a set of estimates, $\widehat{\boldsymbol{\beta}}(m), m = p, .., n$, the corresponding residuals, $\boldsymbol{e}(m) = \boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(m)$, and the robust scales (28.17), $s[\widehat{\boldsymbol{\beta}}(m)]$. The value selected is the $\widehat{\boldsymbol{\beta}}(m)$ which minimizes the robust scale. This process is a complete forward search and several forward searches are done starting with different random samples. The residuals are then identified by using this LMS estimate computed from several forward searches. An improvement of this procedure was proposed by *Atkinson* and *Riani* [28.15], which clearly separates the estimation of the clean subset and the forward search. The initial estimate is computed, as proposed by *Rousseeuw* [28.38], by taking many random samples of size p. The forward search is then applied, but stressing the use of diagnostic statistics to monitor the performance of the procedure.

Finally, *Swallow* and *Kianifard* [28.4] also suggested a similar procedure, which uses a robust estimate of the scale and determines the cutoff values for testing from simulations.

These procedures work when both p and n are not large and the proportion of outliers is moderate, as shown in the simulated comparison by *Wisnowski* et al. [28.39]. However, they do not work as well in large data sets with high contamination. The LMS estimates rely on having at least a sample of size p without outliers, and we need an unfeasible number of samples to have a large probability of this event when p and n are large [28.6]. This good initial estimate is the key for procedures based on clean sets. In the next section we will present procedures that can be applied to large data sets.

28.3.2 Analysis of the Influence Matrix

Let us define the matrix of forecast changes, as the matrix of changes in the forecast of one observation when another observation is deleted. This matrix is given by

$$\mathbf{T} = \begin{pmatrix} \widehat{y}_1 - \widehat{y}_{1(1)} & \widehat{y}_1 - \widehat{y}_{1(2)} & \cdots & \widehat{y}_1 - \widehat{y}_{1(n-1)} & \widehat{y}_1 - \widehat{y}_{1(n)} \\ \widehat{y}_2 - \widehat{y}_{2(1)} & \widehat{y}_2 - \widehat{y}_{2(2)} & \cdots & \widehat{y}_2 - \widehat{y}_{2(n-1)} & \widehat{y}_2 - \widehat{y}_{2(n)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \widehat{y}_{n-1} - \widehat{y}_{n-1(1)} & \widehat{y}_n - \widehat{y}_{n(2)} & \cdots & \widehat{y}_n - \widehat{y}_{n(n-1)} & \widehat{y}_n - \widehat{y}_{n(n)} \\ \widehat{y}_n - \widehat{y}_{n(1)} & \widehat{y}_n - \widehat{y}_{n(2)} & \cdots & \widehat{y}_n - \widehat{y}_{n(n-1)} & \widehat{y}_n - \widehat{y}_{n(n)} \end{pmatrix}.$$

The columns of this matrix are the vectors $\boldsymbol{t}_i = \hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{(i)}$, and Cook's statistic is their standardized norm. These

vectors can also be written as $\boldsymbol{t}_i = \boldsymbol{e}_{(i)} - \boldsymbol{e}$, where $\boldsymbol{e}_{(i)}$ is the vector of residuals when the i-th observation is deleted. Therefore, $\mathbf{T}$ can also be considered the matrix of residual changes. *Peña* and *Yohai* [28.5] define the $n \times n$ influence matrix $\mathbf{M}$ as

$$\mathbf{M} = \frac{1}{ps_{\mathrm{R}}^2}\mathbf{T}'\mathbf{T}.$$

As $\mathbf{H}$ is idempotent it can be shown immediately that $\mathbf{M}$ is given by

$$\mathbf{M} = \frac{1}{ps_{\mathrm{R}}^2}\mathbf{EDHDE}, \tag{28.19}$$

where $\mathbf{E}$ is a diagonal matrix with the residuals on the main diagonal, and $\mathbf{D}$ is a diagonal matrix with elements $(1-h_{ii})^{-1}$. By (28.7) $\mathbf{ED}$ is the diagonal matrix of predictive residuals. Therefore, the ij-th element of $\mathbf{M}$, is

$$m_{ij} = \frac{e_i e_j h_{ij}}{(1-h_{ii})(1-h_{jj})ps_{\mathrm{R}}^2} = \frac{e_{i(i)}e_{j(j)}h_{ij}}{ps_{\mathrm{R}}^2}\,.$$

Assuming that all the residuals are different from zero, from (28.4) the rank of $\mathbf{M}$ is equal to p, the rank of $\mathbf{H}$. Observe that the diagonal elements of $\mathbf{M}$ are the Cook's statistics.

Let $r_{ij} = m_{ij}/m_{ii}^{1/2}m_{jj}^{1/2}$ be the uncentered correlation coefficient between $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$. Let us show that the eigenvectors of the matrix $\mathbf{M}$ will be able to indicate groups of influential observations. Suppose that there are k groups of influential observations $I_1, \ldots, I_k$, such that

1. If $i, j \in I_h$, then $|r_{ij}| = 1$. This means that the effects on the least-squares fit produced by the deletion of two points in the same set I_h have correlation 1 or -1.
2. If $i \in I_j$ and $l \in I_h$ with $j \neq h$, then $r_{il} = 0$. This means that the effects produced on the least-squares fit by observations i and j belonging to different sets are uncorrelated.
3. If i does not belong to any I_h, then $m_{ij} = 0$ for all j. This means that data points outside these groups have no influence on the fit.

Now, according to (1) we can split each set I_h into I_h^1 and I_h^2 such that: (1) if $i, j \in I_h^q$, then $r_{ij} = 1$; (2) if $i \in I_h^1$ and $j \in I_h^2$, then $r_{ij} = -1$. Let $\boldsymbol{v}_1 = (v_{11}, \ldots, v_{1n})', \ldots, \boldsymbol{v}_k = (v_{k1}, \ldots, v_{kn})'$ be defined by $v_{hj} = m_{jj}^{1/2}$ if $j \in I_h^1$; $v_{hj} = m_{jj}^{1/2}$ if $j \in I_h^1$; $v_{hj} = -m_{jj}^{1/2}$ if $j \in I_h^2$ and $v_{hj} = 0$ if $j \notin I_h$. Then, if (1)–(3) hold, by (28.6) the matrix $\mathbf{M}$ is

$$\mathbf{M} = \sum_{i=1}^{k} \boldsymbol{v}_i \boldsymbol{v}_i',$$

and since the $\boldsymbol{v}_i$ are orthogonal, the eigenvectors of $\mathbf{M}$ are $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$, and the corresponding eigenvalues $\lambda_1, \ldots, \lambda_k$ are given by

$$\lambda_h = \sum_{i \in I_h} m_{ii}.$$

It is clear that, when the matrix $\mathbf{M}$ satisfies (1)–(3), the only sets I with large C_I are I_h^q, $1 \leq h \leq k$, $q = 1, 2$, and these sets may be found by looking at the eigenvectors associated with non-null eigenvalues of $\mathbf{M}$. Note that (28.6) can also be written as

$$r_{ij} = \mathrm{sign}(e_i)\mathrm{sign}(e_j)h_{ij}/(h_{ii}h_{jj})^{1/2},$$

which means that, in the extreme case that we have presented, the $\mathbf{H}$ matrix and the signs of the residuals are able, by themselves, to identify the set of points that are associated with masking. For real data sets, (1)–(3) do not hold exactly. However, the masking effect is typically due to the presence of blocks of influential observations in the sample having similar or opposite effects. These blocks are likely to produce a matrix $\mathbf{M}$ with a structure close to that described by (1)–(3). In fact, two influential observations i, j producing similar effects should have r_{ij} close to 1, and close to -1 when they have opposed effects. Influential observations with non-correlated effects have $|r_{ij}|$ close to 0. The same will happen with non-influential observations. Therefore, the eigenvectors will have approximately the structure described above, and the null components will be replaced by small values.

This suggests that we should find the eigenvectors corresponding to the p non-null eigenvalues of the influence matrix $\mathbf{M}$, consider the eigenvectors corresponding to large eigenvalues, and define the sets I_j^1 and I_j^2 by those components with large positive and negative weights, respectively. *Peña* and *Yohai* [28.5] proposed the following procedure.

Step 1: *Identifying sets of outlier candidates*. A set of candidate outlier is obtained by analyzing the eigenvectors corresponding to the non-null eigenvalues of the influence matrix $\mathbf{M}$, and by searching in each eigenvector for a set of coordinates with relatively large weight and the same sign.

Step 2: *Checking for outliers*. (a) Remove all candidate outliers. (b) Use the standard F and t statistics to

Table 28.3 A simulated set of data

	1	2	3	4	5	6	7	8	9(a)	10(a)	9(b)	10(b)	9(c)	10(c)
x	1	2	3	4	5	6	7	8	12	12	12	12	12	12
y	2.0	2.9	3.9	5.1	6.2	6.9	7.8	9.1	19	20	19	7	13	7

Table 28.4 Eigen-analysis of the influence matrix for the data from Table 28.3. The eigenvectors and eigenvalues are shown

	λ_1	λ_1/λ_2	1	2	3	4	5	6	7	8	9	10
(a)	1.27	2.87	−0.17	−0.06	−0.00	−0.00	−0.02	−0.10	−0.22	−0.33	0.42	0.79
(b)	3.78	3.783	0.00	−0.00	−0.00	−0.00	−0.00	0.00	−0.00	−0.00	−0.71	0.71
(c)	3.25	32	−0.05	−0.02	−0.00	−0.00	−0.01	−0.02	−0.04	−0.10	−0.50	0.85

test for groups or individual outliers. Reject sets or individual points with F or t statistics larger than some constant c. For the F statistic the c value corresponds to the distribution of the maximum F over all sets of the same size, and this distribution is unknown. Therefore, it is better to use the t statistic and choose the c value by the Bonferroni inequality or, better still, by simulating the procedure with normal errors. (c) If the number of candidate outliers is larger than $n/2$, the previous procedure can be applied separately to the points identified in each eigenvector.

As an illustration we will use the simulated data from Table 28.3, which are plotted in Fig. 28.1.

The three sets of data have in common cases 1–8 and differ in cases 9 and 10. In the first set of data the largest values of the Cook's statistics are $D_{10} = 0.795$, $D_1 = 0, 29$ and $D_9 = 0.228$. The most influential observation is the 10-th, which has a standardized residual $r_{10} = 1.88$, thus there is no evidence that the point is an outlier. However, the first eigenvector of the influence matrix leads to the results shown in Table 28.4. We see that both cases 9 and 10 appear separated from the rest. When they are deleted from the sample and checked against the first eight observations we obtain the values indicated in Table 28.5, where they are clearly declared as outliers. Thus, in this example the eigenvalues of the influence matrix are able to avoid the masking effect which was clearly present in the univariate statistics.

In case (b), as both outliers have a different sign, they do not produce masking, and both of them are detected by the univariate analysis: $D_9 = 1.889$, and $D_{10} = 1.893$, and the outlier tests are $t_{10} = 5.20$ and $t_9 = -5.24$. The two points are also shown in the extremes of the eigenvalue. Finally in case (c) there is only one outlier which is detected by both the univariate and multivariate analysis.

The influence matrix $\mathbf{M}$ may be considered a generalization of Cook's local influence matrix $\mathbf{L} = \mathbf{EHE}$ [28.17]. It replaces the matrix of residuals $\mathbf{E}$ by the matrix of standardized residuals $\mathbf{ED}$. If there are no high-leverage observations and the h_{ii} are similar for all points, both matrices will also be similar, and will have similar eigenvectors. However, when the observations have very different leverages, the directions corresponding to the eigenvectors of the matrix $\mathbf{M}$ give more weight to the influence of the high-leverage observations, which are

Table 28.5 Values of the t statistics for testing each point as an outlier

Case	9	10
(a)	27.69	32.28
(b)	31.94	−32.09
(c)	−0.07	−32.09

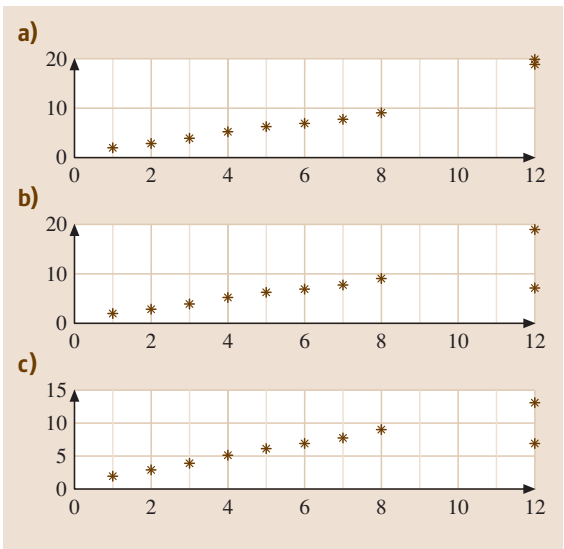

Fig. 28.1 The simulated data from Table 28.3

Table 28.6 Eigenvalues of the sensitivity matrix for the data from Table 28.3

	1	2	3	4	5	6	7	8	9	10
v_1	0.502	0.455	0.407	0.360	0.312	0.264	0.217	0.170	−0.020	−0.020
v_2	−0.191	−0.119	−0.046	0.026	0.099	0.172	0.245	0.318	0.610	0.610

precisely those that are more likely to produce masking effects.

Note that the rank of the influence matrix $\mathbf{M}$ is p, the same as the rank of $\mathbf{H}$, and therefore we do not need to compute n eigenvectors as we only have p eigenvalues linked to nonzero eigenvalues. Thus, the procedure can be applied for very large data sets, see *Peña* and *Yohai* [28.5] for the details of the implementation.

28.3.3 The Sensitivity Matrix

If instead of looking at the columns of the matrix of forecast changes $\mathbf{T}$ we look at its rows, a different perspective appears. The rows indicate the sensitivity of each point, that is, how the forecast for a given point changes when we use as the sample the n sets of $n-1$ data built by deleting each point of the sample. In this way we analyze the sensitivity of a given point under a set of small perturbations of the sample. Let

$$\boldsymbol{s}_i = (\widehat{y}_i - \widehat{y}_{i(1)}, \ldots, \widehat{y}_i - \widehat{y}_{i(n)})'$$

be the i-th row of the matrix $\mathbf{T}$. From (28.14) we can write

$$\boldsymbol{s}_i = (h_{i1}e_1/(1-h_{11}), \ldots, h_{in}e_n/(1-h_{nn})) = \mathbf{ED}\boldsymbol{h}_i,$$

where $\mathbf{E}$ and $\mathbf{D}$ are diagonal matrices of residuals and inverse leverage, respectively, defined in the previous section, and $\boldsymbol{h}_i$ is the i-th column of $\mathbf{H}$. We define the sensitivity matrix by

$$\mathbf{P} = \frac{1}{p\widehat{s}_{\mathrm{R}}^2}\begin{pmatrix} \boldsymbol{s}_1'\boldsymbol{s}_1 & \ldots & \boldsymbol{s}_1'\boldsymbol{s}_n \\ \ldots & \ldots & \ldots \\ \boldsymbol{s}_1'\boldsymbol{s}_n & \ldots & \boldsymbol{s}_n'\boldsymbol{s}_n \end{pmatrix},$$

which can be computed by

$$\mathbf{P} = \frac{1}{p\widehat{s}_{\mathrm{R}}^2}\mathbf{HED}^2\mathbf{EH}, \tag{28.20}$$

and has elements

$$p_{ij} = \frac{1}{p\widehat{s}_{\mathrm{R}}^2}\sum_{k=1}^{n}\frac{e_k^2}{(1-h_{kk})^2}h_{ik}h_{jk}.$$

It can be shown that the sensitivity and the influence matrix have the same eigenvalues and we can obtain the eigenvectors of one matrix from the eigenvectors of the other. *Peña* and *Yohai* [28.6] have shown that eigenvectors of the sensitivity matrix are more powerful for the identification of groups of outliers than those of the influence matrix, although they often lead to the same results. These authors also show that these methods work very well for large sets with many predictors and high levels of contamination.

In the following example we show the use of this matrix for detecting groups of outliers. If we compute the eigenvectors of the sensitivity matrix for the data in Table 28.3 we obtain the results presented in Table 28.6. The first eigenvector clearly separates the observations 9 and 10 from the rest. In fact, if we order the coordinates of this vector we find the largest ratio at $170/20 = 8.5$ which separates cases 9 and 10 from the others.

28.4 A Statistic for Sensitivity for Large Data Sets

The analysis of the eigenvalues of the sensitivity matrix is a very powerful method for finding outliers. However, for large data sets it would be very convenient to have a simple statistic, fast to compute, which can be incorporated into the standard output of regression fitting and which could indicate groups of high-leverage outliers, which are the most difficult to identify. This statistic can be obtained through a proper standardization of the diagonal elements of the sensitivity matrix. *Peña* [28.7] defines the sensitivity statistic at the i-th observation S_i as the squared norm of the standardized vector $\boldsymbol{s}_i$, that is,

$$S_i = \frac{\boldsymbol{s}_i'\boldsymbol{s}_i}{p\widehat{Var}(\hat{y}_i)}, \tag{28.21}$$

and using (28.14) and $\widehat{Var}(\hat{y}_i) = \hat{s}_R^2 h_{ii}$, this statistic can be written as

$$S_i = \frac{1}{p\hat{s}_R^2 h_{ii}} \sum_{j=1}^{n} \frac{h_{ji}^2 e_j^2}{(1-h_{jj})^2}. \tag{28.22}$$

An alternative way to write S_i, is as a linear combination of the sample Cook's distance. From (28.12) and (28.22), we have

$$S_i = \sum_{j=1}^{n} \rho_{ji}^2 D_j, \tag{28.23}$$

where $\rho_{ij} = (h_{ij}^2/h_{ii}h_{jj})^{1/2} \leq 1$ is the correlation between forecasts $\hat{y}_i$ and $\hat{y}_j$. Also, using the predictive residuals, $e_{j(j)} = e_j/(1-h_{jj})$, we have that

$$S_i = \frac{1}{p\hat{s}_R^2} \sum_{j=1}^{n} w_{ji} e_{j(j)}^2 \tag{28.24}$$

and S_i is a weighted combination of the predictive residuals.

The sensitivity statistics has three interesting properties. The first is that, in a sample without outliers or high-leverage observations, all the cases have the same expected sensitivity, approximately equal to $1/p$. This is an important advantage over Cook's statistic, which has an expected value that depends heavily on the leverage of the case. The second property is that, for large sample sizes with many predictors, the distribution of the S_i statistic will be approximately normal. This again is an important difference from Cook's distance, which has a complicated asymptotic distribution [28.37]. This normal distribution allows the computation of cutoff values for finding outliers. The third property is that, when the sample is contaminated by a group of similar outliers with high leverage, the sensitivity statistic will discriminate between the outliers and the good points, and the sensitivity statistic S_i is expected to be smaller for the outliers than for the good data points.

These properties are proved in *Peña* [28.7]. The normality of the distribution of the S_i statistic implies that we can search for outliers by finding observations with large values of $[S_i - E(S_i)]/std(S_i)$. As the possible presence of outliers and high leverage points will affect the distribution of S_i, it is better to use robust estimates such as as the median or the median of the absolute deviations (MAD) from the sample median, and consider as heterogeneous observations those which satisfy:

$$|S_i - med(S)| \geq 4.5\text{MAD}(S_i) \tag{28.25}$$

where $med(S)$ is the median of the S_i values and $\text{MAD}(S_i) = med\,|S_i - med(S)|$. For normal data $\text{MAD}(S_i)/.645$ is a robust estimate for the standard deviation and the previous rule is roughly equivalent to taking three standard deviations in the normal case. In *Peña* [28.7] it is shown that this statistic can be very useful for the diagnostic analysis of large data sets.

28.5 An Example: The Boston Housing Data

As an example of the usefulness of the sensitivity statistics and to compare it with the procedures based on eigenvalues, we will use the Boston housing data set which consists of 506 observations on 14 variables, available at Carnegie Mellon University, Department of Statistics, Pittsburgh (http://lib.stat.cmu.edu). This data set was given by *Belsley* et al. [28.9] and we have used the same variables they considered: the dependent variable is the logarithm of the median value of owner-occupied homes.

Figure 28.2 shows the diagnostic analysis of this data set. The first row corresponds to the residuals of the regression model. The residuals have been divided by their standard error and the first plot shows a few points which can be considered as outliers. The plot of the Studentized residual is similar and identifies the same points as outliers. The second row gives information about Cook's D statistics. There are clearly some points in the middle of the sample which are more influential than the rest, but all the values of the statistic are small and, as we expect a skewed distribution, the conclusion is not clear. However, the sensitivity statistics clearly identifies a group of extreme observations which are not homogeneous with the rest. The median of the sensitivity statistic is 0.0762, which is very close to the expected value $1/p = 1/14 = 0.0714$. The MAD is 0.0195 and the plot indicates that 45 observations are heterogeneous with respect to the rest. These observations are most of the cases 366–425 and some other isolated points. From *Belsley* et al. [28.9] we obtain that cases 357–488 correspond to Boston, whereas the rest correspond to the suburbs. Also, the 45 points indicated by the statistic S_i as outliers all correspond to some central districts of Boston, including the downtown area, which suggests that the relation among the variables could be different in these dis-

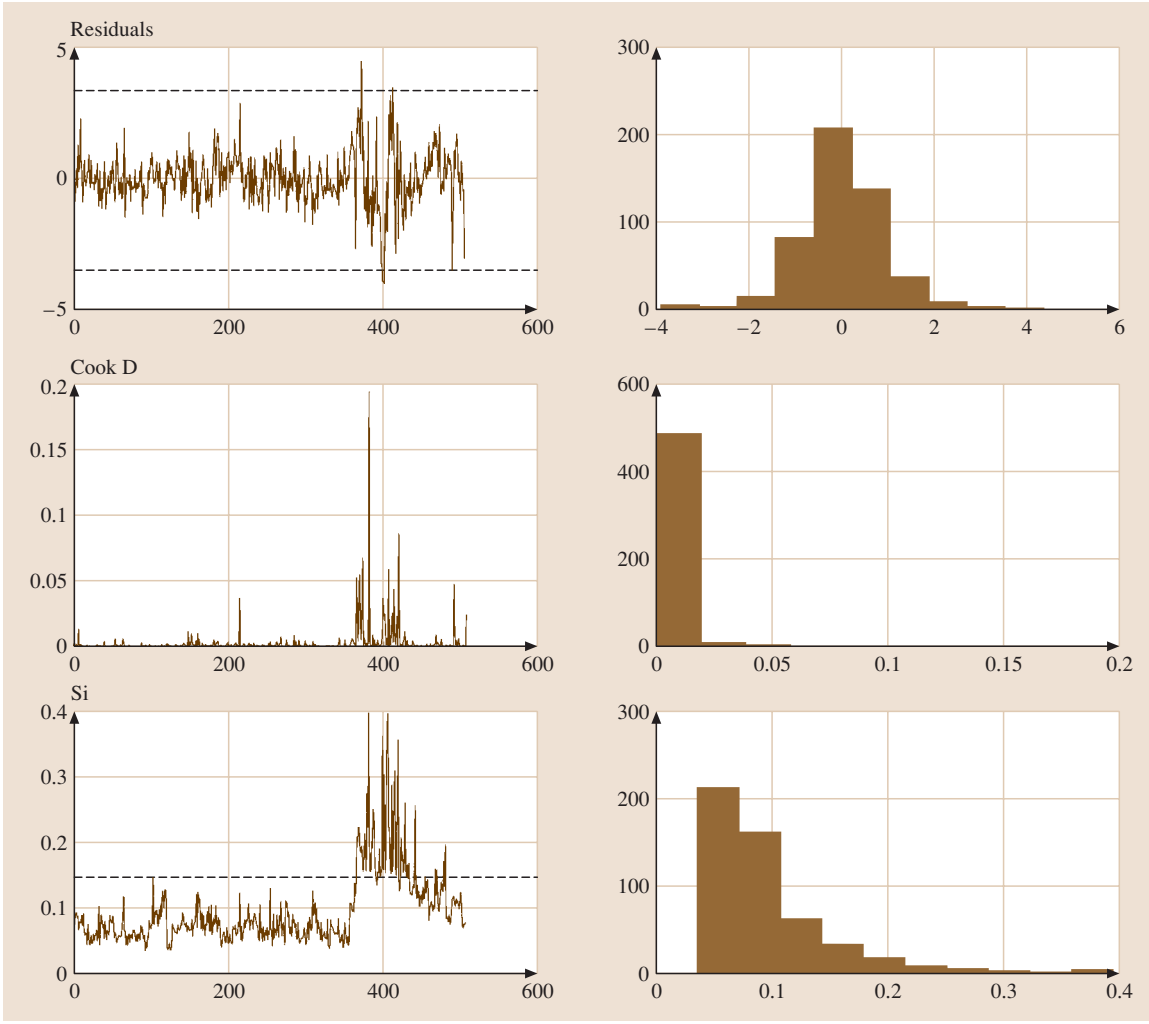

Fig. 28.2 Residuals, Cook's statistics and sensitivity statistics for the Boston housing data. *Right*, histogram; *left* case plot of the value of the statistic

tricts than in the rest of the sample. In fact, if we fit regression equations to these two groups we find very different coefficients for the regression coefficients in both groups of data, and in the second group only five variables are significant. Also, we obtain a large reduction in the residual sum of squares (RSE) when fitting different regression equations in the two groups.

Figure 28.3 shows the first eigenvalues of the matrix of influence and sensitivity. Although both eigenvectors indicate heterogeneity, the one from the matrix of sensitivity is more clear.

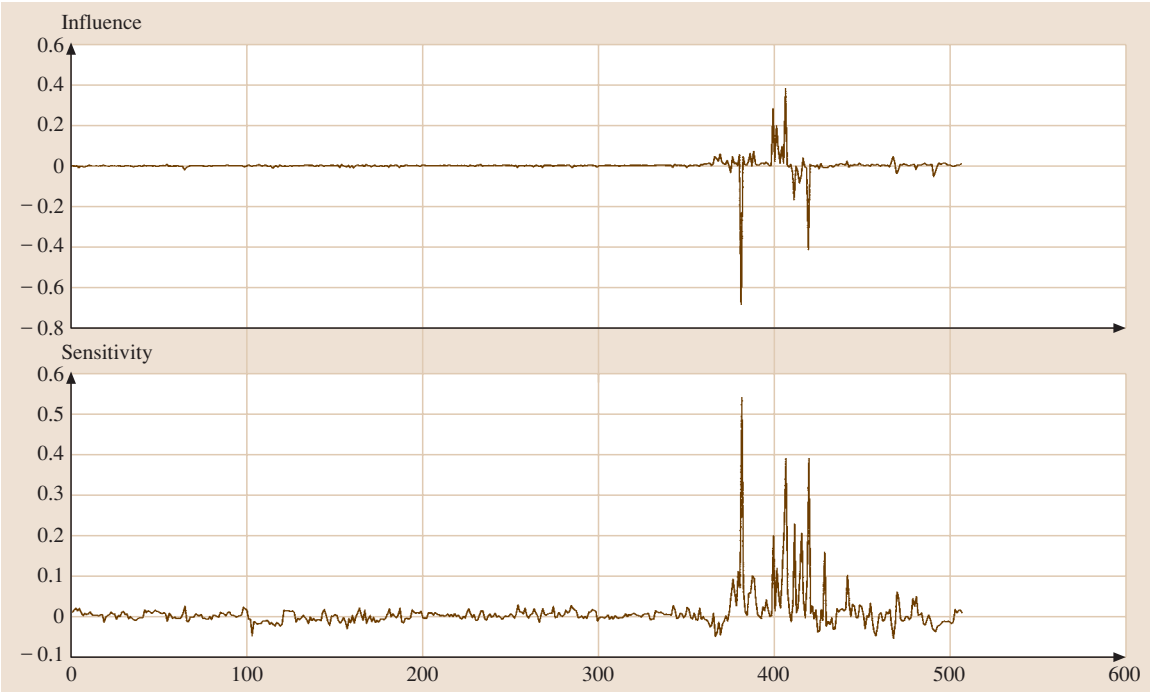

Fig. 28.3 First eigenvalue of the influence and sensitivity matrices

28.6 Final Remarks

We have shown different procedures for diagnosis in regression models and have stressed that the detection of groups of outliers in regression in large data sets can be made by eigen-analysis of the influence and sensitivity matrices. We have also shown that a single statistic of sensitivity is able to reveal masked outliers in many difficult situations. The most challenging problem today is to identify heterogeneity when we do not have a central model which explains more than 50% of the data and groups of outliers, as has been assumed in this article, but different regression models in different regions of the parameter space. In this case robust methods are no longer useful and we need other methods to solve this problem. A promising approach is the split and recombine (SAR) procedure, which has been applied to find heterogeneity in regression models by *Peña* et al. [28.40]. These situations are very close to cluster analysis and finding clusters around different regression lines is today a promising line of research.

References

28.1 A. S. Hadi, J. S. Simonoff: Procedures for the identification of multiple outliers in linear models, J. Am. Statist. Assoc. **88**, 1264–1272 (1993)

28.2 A. S. Hadi, J. S. Simonoff: Improving the estimation and outlier identification properties of the least median of squares and minimum volume ellipsoid estimators, Parisankhyan Samikkha **1**, 61–70 (1994)

28.3 A. C. Atkinson: Fast very robust methods for the detection of multiple outliers, J. Am. Statist. Assoc. **89**, 1329–1339 (1994)

28.4 W. Swallow, F. Kianifard: Using robust scale estimates in detecting multiple outliers in linear regression, Biometrics **52**, 545–556 (1996)

28.5 D. Peña, V. J. Yohai: The detection of influential subsets in linear regression using an influence matrix, J. R. Statist. Soc. B **57**, 145–156 (1995)

28.6 D. Peña, V.J. Yohai: A fast procedure for robust estimation and diagnostics in large regression problems, J. Am. Statist. Assoc. **94**, 434–445 (1999)

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

28.8 G.E.P. Box: When Murphy speaks listen, Qual. Prog. **22**, 79–84 (1989)

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

28.10 D.M. Hawkins: *Identification of Outliers* (Chapman Hall, New York 1980)

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

28.12 A.C. Atkinson: *Plots, Transformations and Regression* (Clarendon, Oxford 1985)

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

28.14 V. Barnett, T. Lewis: *Outliers in Statistical Data*, 3 edn. (Wiley, New York 1994)

28.15 A.C. Atkinson, M. Riani: *Robust Diagnostic Regression Analysis* (Springer, Berlin Heidelberg New York 2000)

28.16 R.D. Cook: Detection of influential observations in linear regression, Technometrics **19**, 15–18 (1977)

28.17 R.D. Cook: Assessment of local influence (with discussion), J. R. Statist. Soc. B **48**(2), 133–169 (1986)

28.18 G.C. Brown, A.J. Lawrence: Theory and ilustration of regression influence diagnostics, Commun. Statist. A **29**, 2079–2107 (2000)

28.19 M. Suárez Rancel, M.A. González Sierra: Regression diagnostic using local influence: A review, Commun. Statist. A **30**, 799–813 (2001)

28.20 G. Hartless, J. G. Booth, R.C. Littell: Local influence of predictors in multiple linear regression, Technometrics **45**, 326–332 (2003)

28.21 F. Critchley, R.A. Atkinson, G. Lu, E. Biazi: Influence analysis based on the case sensitivity function, J. R. Statist. Soc. B **63**(2), 307–323 (2001)

28.22 J. Lawrance: Deletion influence and masking in regression, J. R. Statist. Soc. B **57**, 181–189 (1995)

28.23 D.M. Hawkins, D. Bradu, G.V. Kass: Location of several oultiers in multiple regression data using elemental sets, Technometrics **26**, 197–208 (1984)

28.24 J.B. Gray, R.F. Ling: K–Clustering as a detection tool for influential subsets in regression, Technometrics **26**, 305–330 (1984)

28.25 M.G. Marasinghe: A multistage procedure for detecting several outliers in linear regression, Technometrics **27**, 395–399 (1985)

28.26 F. Kianifard, W. Swallow: Using recursive residuals calculated in adaptively ordered observations to identify outliers in linear regression, Biometrics **45**, 571–585 (1989)

28.27 F. Kianifard, W. Swallow: A Monte Carlo Comparison of five Procedures for Identifying Outliers in Lineal Regression, Commun. Statist. (Theory and Methods) **19**, 1913–1938 (1990)

28.28 A.C. Atkinson: Masking unmasked, Biometrika **73**, 533–41 (1986)

28.29 P. Huber: Between Robustness and Diagnosis. In: *Directions in Robust Statistics and Diagnosis*, ed. by W. Stahel, S. Weisberg (Springer, Berlin Heidelberg New York 1991) pp. 121–130

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

28.31 R.A. Maronna, R.D. Martin, V.J. Yohai: *Robust Statistics, Theory and Practice* (Wiley, New York 2006)

28.32 G.E.P. Box, C.G. Tiao: A Bayesian approach to some outlier problems, Biometrika **55**, 119–129 (1968)

28.33 A. Justel, D. Peña: Bayesian unmasking in linear models, Comput. Statist. Data Anal. **36**, 69–94 (2001)

28.34 D. Peña, I. Guttman: Comparing probabilistic models for outlier detection, Biometrika **80**(3), 603–610 (1993)

28.35 J.A. Diaz-Garcia, G. Gonzalez-Farias: A note on the Cook's distance, J. Statist. Planning Inference **120**, 119–136 (2004)

28.36 R.D. Cook, D. Peña, S. Weisberg: The likelihood displacement. A unifying principle for influence, Commun. Statist. A **17**, 623–640 (1988)

28.37 E.K. Muller, M.C. Mok: The disribution of Cook's D statistics, Commun. Statist. A **26**, 525–546 (1997)

28.38 P.J. Rousseeuw: Least median of squares regression, J. Am. Statist. Assoc. **79**, 871–880 (1984)

28.39 J.W. Wisnowski, D.C. Montgomey, J.R. Simpson: A comparative analysis of multiple outliers detection procedures in the linear regression model, Comput. Statist. Data Anal. **36**, 351–382 (2001)

28.40 D. Peña, J. Rodriguez, G.C. Tiao: Identifying mixtures of regression equations by the SAR procedure (with discussion). In: *Bayesian Statistics*, Vol. 7, ed. by Bernardo et al. (Oxford Univ. Press, Oxford 2003) pp. 327–347