Stevens, J. (1984). Outliers & influential datapoints in regression analysis. Psychological Bulletin, 95, 334-344

Outliers and Influential Data Points in Regression Analysis

James P. Stevens
University of Cincinnati

Because the results of a regression analysis can be quite sensitive to outliers (either on y or in the space of the predictors), it is important to be able to detect such points. This article discusses and interrelates the following four diagnostics that are useful in identifying outliers: studentized residuals, the hat elements, Cook's distance, and Mahalanobis distance. Guidelines are given for interpretation of the diagnostics. Outliers will not necessarily be influential in affecting the regression coefficients. This important fact is illustrated and emphasized.

An important area that has been treated only briefly by regression texts that psychologists would typically use (Cohen & Cohen, 1975; Pedhazur, 1982) is that of outliers and influential data points. Because of this lack of research many applied investigators are unaware of how sensitive multiple regression (and other mathematical maximization procedures such as canonical correlation, discriminant analysis, and principal components analysis) can be to just 1 or 2 errant points. These points can substantially affect the results and the subsequent interpretation. It is certainly moot as to whether 1 or 2 points should have such a profound influence. Therefore, it is important to be able to detect outliers and influential points. There is a distinction between the two because a point that is an outlier (either on y or in the space of the predictors) will not necessarily be influential in affecting the regression equation. Examples are given later to illustrate this concept.

Since the late 1970s, statisticians have given much consideration to "case analysis," in which various statistics are used to detect outliers and high influence points (Cook, 1977; Hoaglin & Welsch, 1978). The importance of exploring the data, rather than just examining one or two summary statistics, has been emphasized by Tukey (1977) and many other statisticians. Case analysis in regression is simply a reflection of this emphasis.

The author wishes to thank the associate editor and an anonymous reviewer for their helpful comments.

Requests for reprints should be sent to James P. Stevens, Teachers College Building, University of Cincinnati, Cincinnati, Ohio 45221.

The fact that a simple examination of summary statistics can result in misleading interpretations was illustrated by Anscombe (1973). He presented three data sets that yielded the same summary statistics (i.e., regression coefficients and same $r^2 = .667$). In one case linear regression was perfectly appropriate. In the second case, however, a scatter plot showed that curvilinear regression was appropriate. In the third case, linear regression was appropriate for 10 of 11 points, but the other point was an outlier and possibly should have been excluded from the analysis.

The need to consider the influence of each case on a simple correlation, through the use of influence curves, has been emphasized by Thissen, Baker, & Wainer (1981). This article extends the discussion to the multiple regression context and considers and interrelates statistics useful in detecting errant points. These statistics include Cook's distance (for high influence points), the studentized residuals (for outliers on y), and the Mahalanobis distance and the diagonal elements of the hat matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (for outliers in the space of the predictors). These important diagnostic statistics recently became available on widely distributed statistical software packages: SAS (Barr, Goodnight, Sall, & Heboig, 1982), BMED (Dixon, 1981), and Statistical Package for the Social Sciences (SPSS; Hull & Nie, 1981). This study illustrates the use of these statistics on some small data sets and gives guidelines for their interpretation. Throughout the article, reference is made to three major statistical packages (BMDP, SPSS, and SAS) because, in practice, the vast majority of researchers will be using one or more of these for their analyses.

2, 334-344
Association, Inc.

There are two basic approaches that can be used in dealing with outliers and influential points. This report considers the approach of having an arsenal of tools for isolating these important points for further study, with the possibility of deleting some or all of the points from the analysis. The other approach is to develop procedures that are relatively insensitive to wild points (i.e., robust regression techniques). Some pertinent references for robust regression are (Hogg, 1979; Huber, 1977; Mosteller & Tukey, 1977). It is important to note that even robust regression may be ineffective when there are outliers in the space of the predictors (Huber, 1977). Thus, even in robust regression there is a need for case analysis. Also, a modification of robust regression, called bounded-influence regression, has been developed by Krasker and Welsch (1979).

Data Editing

Influential cases can occur because of recording errors. Consequently, researchers should give more consideration to the data editing phase of the data analysis process (i.e., always listing the data and examining the list for possible errors). There are many possible sources of error from the initial data collection to the final keypunching. First, some of the data may have been recorded incorrectly. Second, even if recorded correctly, when all of the data are transferred to a single sheet or a few sheets in preparation for keypunching, errors may be made. Finally, even if no errors are made in these first two steps, an error could be made in putting the data on cards or into the terminal.

Graphical Diagnostic Procedures

Some ways in which traditional graphical plots can be quite useful are examined (cf. Draper & Smith, 1981, chap. 3 for an extended discussion). Other ways in which plots are limited are then considered. Plots can be a useful device for detecting outliers, especially when the number of predictors is small, and for diagnosing various reasons for model failure (i.e., nonlinearity, nonconstant variance). Univariate distribution plots for each predictor will reveal at least some of the cases that may be outliers in the space of the predictors. A histogram of the studentized residuals will reveal which residuals split off from the rest and therefore correspond to possible outliers on y for those cases. Also, for the two-predictor case, a scatter plot of x_1 versus x_2 will indicate which points split off from the cluster of points in the plane and therefore are multivariate outliers.

Plots of studentized residuals ($r_{(-i)}$ versus y_i and x_i are useful. Equation 4 presents the formula for the studentized residuals. These residuals are essentially uncorrelated with y_i and x_i. Therefore, if the model fit is satisfactory, a plot of $r_{(-i)}$ versus y_i or x_i will show a random scattering of points about the line $r_{(-i)} = 0$. A systematic trend indicates model failure. For example, if the points funnel in (or out) as x_i increases, this indicates nonconstant variance. If the points have a curvilinear clustering when plotting y_i versus $r_{(-i)}$, then nonlinearity is present. See Weisberg (1980, chap. 6) for remedies for these problems.

The use of these simple plots decreases as the number of predictors increases. Multivariate outliers in the space of the predictors can then occur in more subtle ways. For example, there might be fairly large differences on 3 of 5 predictors or moderate differences on 7 of 8 predictors, each of which produces a multivariate outlier. Also, the simple plots do not indicate the extent to which deleting 1 or 2 points from the analysis will affect the regression coefficients or the standard errors. Additionally, although the plots can vividly bring possible outliers to one's attention, these need to be tested analytically to guard against overall Type I errors. This is because chance is being capitalized on in testing the cases with the largest values. This is discussed in more detail in the next section.

Belsley, Kuh, and Welsch (1980, p. 7) summarized this discussion as follows:

> Before performing a multiple regression, it is common practice to look at the univariate distribution of each variate to see if any oddities (outliers or gaps) strike the eye. Scatter diagrams are also examined. While there are clear benefits from sorting out peculiar observations in this way, diagnostics of this type can't detect multivariate discrepant observations, nor can they tell us in what ways such data influence the estimated model.

Measuring Outliers

Studentized Residuals

The raw residuals ($\hat{e}_i = y_i - \hat{y}_i$) in linear regression are assumed to be independent, to have a mean of 0, to have a constant variance, and to follow a normal distribution. However, because the n residuals have only $n - k$ degrees of freedom (k degrees of freedom were lost in estimating the regression parameters), they cannot be independent. Also, the residuals have different variances. It can be shown (cf. Draper & Smith, 1981, p. 151) that the covariance matrix for the residuals is given by

$$\mathbf{V}(e) = \hat{\sigma}^2(\mathbf{I} - \mathbf{H}), \tag{1}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix and $\mathbf{X}$ is on $nx(p + 1)$ matrix with 1s in the first column and the scores for the predictors in the remaining columns. From Equation 1, it is apparent that the variance for the ith residual is given by

$$s_{e_i}^2 = \hat{\sigma}^2(1 - h_{ii}), \tag{2}$$

where h_{ii} is the ith diagonal element of $\mathbf{H}$. This shows clearly that the extent to which the variances of the e_i vary is a direct function of the variability of the diagonal elements of the hat matrix. Equation 2 is more important in describing the role of the h_{ii} as influential points in regression analysis.

In order to compare residuals directly, each e_i is divided by its estimated standard deviation. The resulting residual, r_i, is the standardized residual.

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}. \tag{3}$$

Because the r_i is assumed to have a normal distribution with a mean of 0 (if the model is correct), then about 99% of the r_i should lie within 3 standard deviations of the mean. Therefore, any r_i with an absolute value greater than 3 is an outlier and should be carefully examined to explain why the fit was so poor. The subjects corresponding to these points may be quite different from the rest of the subjects in some important ways.

It is necessary to distinguish between the standardized residual and the studentized residual $r_{(-i)}$, which is defined as

$$r_{(-i)} = \frac{\hat{e}_i}{\hat{\sigma}_{(-i)}\sqrt{1 - h_{ii}}}, \tag{4}$$

where $\hat{\sigma}_{(-i)}$ is the standard error of y with the ith case deleted.

By reexpressing $r_{(-i)}$, the two residuals can easily be related:

$$r_{(-i)} = \frac{\hat{\sigma}\hat{e}_i}{\hat{\sigma}_{(-i)}\hat{\sigma}\sqrt{1 - h_{ii}}} = \frac{\hat{\sigma}}{\hat{\sigma}_{(-i)}} r_i. \tag{5}$$

Because of Equation 5, the studentized residual will be a more sensitive outlier detector; it is therefore preferable. The studentized residual is available in PROC REG (SAS, Barr et al., 1982), whereas BMDP9R gives the standardized residual.[1]

Weisberg (1980) has given the following t statistic for testing an outlier for significance:

$$t_i = r_i\sqrt{\frac{n - p' - 1}{n - p' - r_i^2}}, \tag{6}$$

where r_i is the standardized residual, n is sample size, p' is the number of parameters (including the regression constant), and $df = n - p' - 1$. Although t_i is not given on the printout from BIOMED, SPSS, OR SAS, $r_{(-i)}$ is given by SAS and should be used in Equation 6 instead of r_i. Beckman and Trussel (1974) have shown that $r_{(-i)}$ is also distributed as a t statistic.

Assessing the significance of the case with the largest value of t_i is equivalent to performing n significance tests, one for each of the n cases. To control the resulting inflated Type I error rate, the somewhat conservative Bonferroni inequality is used, doing each test at the α/n level of significance. Table 1 gives the critical values (Weisberg, 1980) for various n and p', which keeps overall $\alpha = .05$.

Example

Consider a regression analysis with 4 predictors on 50 subjects; the largest $r_i = 2.8$. Is this a statistically significant deviation (overall $\alpha = .05$) according to the Weisberg test?

$$t_i = r_i\sqrt{\frac{n - p' - 1}{n - p' - r_i^2}}$$

$$= 2.8\sqrt{\frac{50 - 5 - 1}{50 - 5 - 7.84}}$$

$$= 2.8(1.088) = 3.047.$$

[1] As of June 1983, BMDP2R has several new regression diagnostics available, one of which is the studentized residual (although it is called the jackknived residual).

Because the critical value is 3.53, this is not a significant deviation.

Simply because a case has a larger outlier on y does not necessarily mean the case will be influential in affecting the regression coefficients. An example is taken from Draper and Smith (1981, p. 169). In this example, there are 19 data points and one predictor. The scatter plot is presented in Figure 1. Attention should be focused on the line labeled "Line without point 2." Inspection of Figure 1 reveals that the case labeled 3 is an outlier, but it is not influential because, being "outweighed" by points at neighboring values, it cannot have much effect on the values of the estimated coefficients of the model. The case labeled 1, in contrast, is an outlier on x and is an influential point. In fact, without Case 1, the regression is not significant. This example is discussed in greater detail later. The reader might generalize and erroneously conclude that whenever an observation on the predictor is an outlier it will be influential. This is not necessarily true. To determine whether an outlier (either on y or on the set of predictors) is influential Cook's measure of distance is used.

Measures for Outliers in the Space of the Predictors

Diagonal elements h_{ii} of the hat matrix. The h_{ii}s are one measure of the extent to which the ith observation is an outlier in the space of the predictors. The h_{ii}s are important because they can play a key role in determining the predicted values for the subjects. Recall that

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{and} \quad \hat{\mathbf{y}} = \mathbf{X}\hat{\beta}.$$

Therefore, $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ by simple substitution.

Thus, the predicted values for y are obtained by postmultiplying the hat matrix by the column vector of observed scores on y. It can be shown that the h_{ii}s lie between 0 and 1, and that the average value for $h_{ii} = p/n$. From Equation 2, it may be seen that when h_{ii} is large (i.e., near 1), then the variance for the ith residual is near 0. This means that $y_i \approx \hat{y}_i$. In other words, an observation may fit the linear model well and yet be an influential data point. This second diagnostic, then, is "flagging" observations that need to be examined carefully because they may have an unusually large influence on the regression coefficients.

What is a significant value for the h_{ii}? Hoaglin and Welsch (1978) suggest that $2p/n$ may be considered large. Belsey et al. (1980, pp. 67–68) show that when the set of predictors is multivariate normal, then $(n - p)[h_{ii} - 1/n]/(1 - h_{ii})(p - 1)$ is distributed as F with $(p - 1)$ and $(n - p)$ degrees of freedom. Rough, quick, approximate critical values (for $\alpha = .05$) are available to cover most situations. For $3 < p < 9$ and $25 < n - p < 50$, $3p/n$ can be used, whereas for $p > 10$ and $n - p > 50$, $2p/n$ is appropriate.

An important point to remember concerning the hat elements is that the points they identify will not necessarily be influential in affecting the regression coefficients. This does not mean, however, that their identification served no purpose for, as Draper and John (1981) note, "even if they were not [influential], it would be useful to know that the points were remote. For example, if additional observations were planned, such information would clearly be helpful" (p. 25).

Andrews and Pregibon (1978) have proposed a statistic (AP), which is a generalization of h_{ii} for identifying whether a set of points is remote in the space of the predictors. For one point their statistic reduces to $1 - h_{ii}$; thus, a small value of the AP statistic corresponds to a large value for h_{ii}. The use of the AP statistic is discussed in Draper and John (1981). The AP statistic is actually based on earlier work by Wilks (1963), who used the notion of the determinant of a matrix as the multivariate measure of scatter. Wilks suggested that a reasonable criterion for an outlier is a point i, which, when omitted, would minimize the following ratio:

$$R_{(-i)} = \frac{|\mathbf{X}'\mathbf{X}^{(i)}|}{|\mathbf{X}'\mathbf{X}|},$$

where $|\mathbf{X}'\mathbf{X}^{(-i)}|$ is the determinant of $\mathbf{X}'\mathbf{X}$ with the vector for case i deleted from $\mathbf{X}$. Minimizing R is, of course, equivalent to deleting the point which has the greatest effect on the multivariate dispersion. Wilks generalized $R_{(-i)}$ to test whether a set of s points is remote by considering the previous ratio of determinants, in which the numerator involves deleting the vectors for the s cases from $\mathbf{X}'\mathbf{X}$.

Barnett and Lewis (1978) give critical values for $p = 2$ through 5 predictors for determining

Table 1
Critical Values for Outlier Test With Overall α = .05

n/p'	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30
6	4.85	6.23	10.89	76.39														
7	4.38	5.07	6.58	11.77	89.12													
8	4.12	4.53	5.26	6.90	12.59	101.9												
9	3.95	4.22	4.66	5.44	7.18	13.36	114.6											
10	3.83	4.03	4.32	4.77	5.60	7.45	14.09	127.3										
11	3.75	3.90	4.10	4.40	4.88	5.75	7.70	14.74	140.1									
12	3.69	3.81	3.96	4.17	4.49	4.98	5.89	7.94	15.44	152.8								
13	3.65	3.74	3.86	4.02	4.24	4.56	5.08	6.02	8.16	16.08	165.5							
14	3.61	3.69	3.79	3.91	4.07	4.30	4.63	5.10	6.14	8.37	16.69	178.2						
15	3.58	3.65	3.73	3.83	3.95	4.12	4.36	4.70	5.25	6.25	8.58	17.28	191.0					
16	3.56	3.62	3.68	3.77	3.87	4.00	4.17	4.41	4.76	5.33	6.36	8.77	17.85	203.7				
17	3.54	3.59	3.65	3.72	3.80	3.90	4.04	4.21	4.46	4.82	5.40	6.47	8.95	18.40	216.4			
18	3.53	3.57	3.62	3.68	3.75	3.83	3.94	4.08	4.26	4.51	4.88	5.47	6.57	9.13	18.93			
19	3.52	3.56	3.60	3.65	3.71	3.78	3.86	3.97	4.11	4.30	4.55	4.93	5.54	6.67	9.30			
20	3.51	3.54	3.58	3.62	3.67	3.73	3.81	3.89	4.00	4.15	4.33	4.59	4.98	5.60	6.76			
21	3.50	3.53	3.57	3.60	3.65	3.70	3.76	3.83	3.92	4.03	4.18	4.37	4.64	5.03	5.67			
22	3.50	3.52	3.55	3.59	3.63	3.67	3.72	3.78	3.86	3.95	4.06	4.21	4.40	4.68	5.08	280.1		
23	3.49	3.52	3.54	3.57	3.61	3.65	3.69	3.75	3.81	3.88	3.98	4.09	4.24	4.44	4.71	21.41		
24	3.49	3.51	3.53	3.56	3.59	3.63	3.67	3.71	3.77	3.83	3.91	4.00	4.12	4.27	4.47	10.07		
25	3.48	3.50	3.53	3.55	3.58	3.61	3.65	3.69	3.73	3.79	3.85	3.93	4.02	4.14	4.30	7.17		
26	3.48	3.50	3.52	3.54	3.57	3.60	3.63	3.66	3.70	3.75	3.81	3.87	3.95	4.05	4.17	5.95		
27	3.48	3.50	3.52	3.54	3.56	3.58	3.61	3.65	3.68	3.72	3.77	3.83	3.89	3.97	4.07	5.29	343.8	
28	3.48	3.50	3.51	3.53	3.55	3.58	3.60	3.63	3.66	3.70	3.74	3.79	3.84	3.91	3.99	4.88	23.63	
29	3.48	3.49	3.51	3.53	3.55	3.57	3.59	3.62	3.64	3.68	3.71	3.76	3.81	3.86	3.93	4.61	10.74	
30	3.48	3.49	3.51	3.52	3.54	3.56	3.58	3.60	3.63	3.66	3.69	3.73	3.77	3.82	3.88	4.42	7.53	
31	3.48	3.49	3.50	3.52	3.54	3.55	3.57	3.59	3.62	3.64	3.67	3.71	3.74	3.79	3.84	4.28	6.18	
32	3.48	3.49	3.50	3.52	3.53	3.55	3.57	3.59	3.61	3.63	3.66	3.69	3.72	3.76	3.80	4.17	5.47	407.4
33	3.48	3.49	3.50	3.52	3.53	3.54	3.56	3.58	3.60	3.62	3.64	3.67	3.70	3.74	3.77	4.08	5.03	25.66
34	3.48	3.49	3.50	3.51	3.53	3.54	3.56	3.57	3.59	3.61	3.63	3.66	3.68	3.71	3.75	4.01	4.74	11.34
35	3.48	3.49	3.50	3.51	3.52	3.54	3.55	3.57	3.58	3.60	3.62	3.64	3.67	3.70	3.73	3.96	4.53	7.84
36	3.48	3.49	3.50	3.51	3.52	3.54	3.55	3.56	3.58	3.60	3.61	3.63	3.66	3.68	3.71	3.91	4.37	6.39
37	3.48	3.49	3.50	3.51	3.52	3.53	3.55	3.56	3.57	3.59	3.61	3.62	3.65	3.67	3.69	3.87	4.26	5.62
38	3.48	3.49	3.50	3.51	3.52	3.53	3.54	3.56	3.57	3.58	3.60	3.62	3.64	3.66	3.68	3.84	4.16	5.16
39	3.49	3.49	3.50	3.51	3.52	3.53	3.54	3.55	3.57	3.58	3.59	3.61	3.63	3.65	3.67	3.81	4.09	4.84
40	3.49	3.49	3.50	3.51	3.52	3.53	3.54	3.55	3.57	3.58	3.59	3.60	3.62	3.64	3.66	3.79	4.03	4.62
50	3.51	3.51	3.51	3.52	3.53	3.53	3.54	3.54	3.55	3.56	3.57	3.57	3.58	3.59	3.60	3.66	3.75	3.88

Table 1 (*continued*)

n/p'	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20	25	30
60	3.53	3.53	3.53	3.54	3.54	3.54	3.55	3.55	3.56	3.56	3.57	3.57	3.58	3.58	3.59	3.62	3.67	3.73
70	3.55	3.55	3.55	3.55	3.56	3.56	3.56	3.56	3.57	3.57	3.57	3.58	3.58	3.59	3.59	3.61	3.64	3.67
80	3.57	3.57	3.57	3.57	3.57	3.58	3.58	3.58	3.58	3.58	3.59	3.59	3.59	3.60	3.60	3.61	3.63	3.66
90	3.58	3.59	3.59	3.59	3.59	3.59	3.59	3.60	3.60	3.60	3.60	3.60	3.60	3.61	3.61	3.62	3.63	3.65
100	3.60	3.60	3.60	3.60	3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.62	3.62	3.62	3.62	3.63	3.64	3.65
200	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.74	3.74	3.74	3.74
300	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.82	3.82	3.82	3.82	3.82	3.82	3.82	3.82
400	3.87	3.87	3.87	3.87	3.87	3.87	3.87	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88
500	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92

Note. From *Applied Linear Regression* by S. Weisberg, 1980, New York: Wiley. Copyright 1980 by Wiley. Reprinted by permission of John Wiley & Sons, Ltd.

which pairs of points are significantly separated. Detection of such sets for higher dimensional p will be far from obvious. Also, computation for all pairs, triples, and so on, if n is even greater than 20, will be quite excessive. An example illustrating the detection of influential pairs and triples, when the goal is to predict future observations, is provided by W. Johnson and Geisser (1983).

Mahalanobis distance. This gives the distance from the case to the centroid of all cases for the predictor variables. A large distance indicates an observation that is an outlier in the space defined by the predictors. The Mahalanobis distance can be written in terms of the covariance matrix **S** as

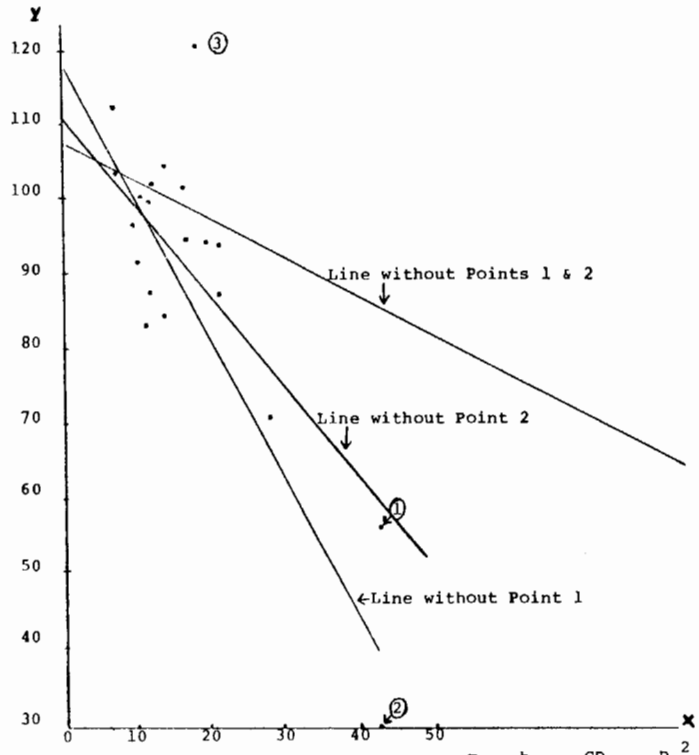
$$D_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})'\mathbf{S}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}})', \qquad (7)$$

where $\mathbf{X}_i$ is the vector of the data for case i and $\bar{\mathbf{X}}$ is the vector of means (centroid) for the predictors.

For a better understanding of D_i^2, consider two small data sets. The first set has two predictors. In Table 2, the data is presented, as well as the D_i^2 and the descriptive statistics (including **S**). The D_i^2 for Cases 6 and 10 are large because the score for Case 6 on x_1 (150) was deviant, whereas for Case 10 the score on x_2 (97) was very deviant. The graphical split-off of Cases 6 and 10 is quite vivid and is displayed in Figure 2.

In the previous example, because the numbers of predictors and subjects were few, it would have been fairly easy to spot the outliers even without the Mahalanobis distance. However, in practical problems with 200 or 300 subjects and 10 predictors, outliers are not always easy to spot and can occur in more subtle ways. For example, a case may have a large distance because there are moderate to fairly large differences on many of the predictors. The second small data set with 4 predictors and $N = 15$ in Table 2 illustrates this latter point. The D_i^2 for case 13 is quite large (7.97) even though the scores for that subject do not split off in a striking fashion for any of the predictors. Rather, it is a cumulative effect that produces the separation.

How large must D_i^2 be before one can say that case i is significantly separated from the rest of the data at the .05 level of significance? If it is tenable that the predictors came from a multivariate normal population, then the



Line without	Regression coeffs.	r_i	h_{ii}	CD_i	D_i^2
Point 2	111.51, -1.18	-.753	.42	.508	10.604
Point 1	118.64, -1.76	-2.00	.42	3.583	10.604
Points 1 & 2	107.92, -.89				

Figure 1. Regression lines and diagnostics for Mickey, Dunn, & Clark (1967) data. (A variation (42.30) on data point 1 has been added.)

critical values (Barnett & Lewis, 1978) are given in Table 3 for 2 through 5 predictors. An easily implemented graphical test for multivariate normality is available (R. Johnson & Wichern, 1982). The test involves plotting ordered Mahalanobis distances against chi-square percentile points. The D_i^2 can be obtained from the BMDP9R.

Referring back to the example with 2 predictors and $n = 10$, if we assume multivariate normality, then Case 6 ($D_i^2 = 5.48$) is not significantly separated from the rest of the data at .05 level because the critical value equals 6.32. In contrast, Case 10 is significantly separated.

BMDP and SPSS give the Mahalanobis distance, whereas SAS gives the h_{ii} (hat elements). It might appear that these are yielding somewhat different information. However, in general, this is not so as seen in the following (Weisberg, 1980, p. 104):

$$h_{ii} = 1/n + (\mathbf{X}_i - \bar{\mathbf{X}})'\mathbf{C}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}). \quad (8)$$

If n is even moderately large (50 or more), then $1/n$ is very small. When the remaining portion is multiplied by $n - 1$, the result is D_i^2 (Mahalanobis distance). In other words, for large n, D_i^2 is approximately proportional to h_{ii}:

$$D_i^2 \approx (n - 1)h_{ii}. \quad (9)$$

Thus, with large n, either measure may be used. Also, because this report previously indicated what would correspond roughly to a

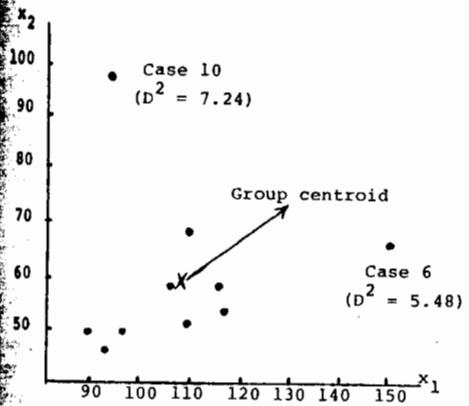


Figure 2. Plot of data for two-predictor example and Mahalanobis distances for outliers.

significant h_{ii} value, from Equation 9 we can immediately determine the corresponding significant D_i^2 value. For example, if $p = 7$ and $n = 50$, then a large $h_{ii} = .42$ and the corresponding large $D_i^2 = 20.58$. If $p = 20$ and $n = 200$, then a large $h_{ii} = 2p/n = .20$ and the corresponding large $D_i^2 = 39.90$.

A Measure for Influential Data Points

Cook's distance. Cook's distance (CD_i) is a measure of the change in the regression coefficients that would occur if this case was omitted, thus revealing which cases are most influential in affecting the regression equation. It is affected by both the case being an outlier on y and on the set of the predictors. Cook's distance is given by

$$CD_i = (\hat{\beta} - \hat{\beta}_{(-i)})'\mathbf{X}'\mathbf{X}(\hat{\beta} - \hat{\beta}_{(-i)})/(p + 1)MS_{res}, \quad (10)$$

where $\hat{\beta}_{(-i)}$ is the vector of estimated regression coefficients with the ith data point deleted, p is the number of predictors, and MS_{res} is the residual (error) variance for the full data set.

Removing the ith data point should keep

Table 2
Raw Data and Mahalanobis Distances for Two Small Data Sets

Case	Y	X_1	X_2	X_3	X_4	D_i^2
1	476	111	68	17	81	0.30
2	457	92	46	28	67	1.55
3	540	90	50	19	83	1.47
4	551	107	59	25	71	0.01
5	575	98	50	13	92	0.76
6	698	150	66	20	90	5.48
7	545	118	54	11	101	0.47
8	574	110	51	26	82	0.38
9	645	117	59	18	87	0.23
10	556	94	97	12	69	7.24
11	634	130	57	16	97	
12	637	118	51	19	78	
13	390	91	44	14	64	
14	562	118	61	20	103	
15	560	109	66	13	88	
Summary statistics						
M	561.70000	108.70000	60.00000			
SD	70.74846	17.73289	14.84737			
S =		$\begin{bmatrix} 314.455 & 19.483 \\ 19.483 & 220.444 \end{bmatrix}$				

Note. Boxed-in entries are the first data set and corresponding D_i^2. The 10 case numbers having the largest D_i^2 for a four-predictor data set are: 10, 10.859; 13, 7.977; 6, 7.223; 2, 5.048; 14, 4.874; 7, 3.514; 5, 3.177; 3, 2.616; 8, 2.561; 4, 2.404.

Calculation of D_i^2 for Case 6:

$$D_6^2 = (41.3, 6) \begin{bmatrix} 314.455 & 19.483 \\ 19.483 & 220.44 \end{bmatrix}^{-1} \begin{pmatrix} 41.3 \\ 6 \end{pmatrix}$$

$$S^{-1} = \begin{matrix} .0032 & -.00029 \\ -.00028 & .00456 \end{matrix} \rightarrow D_6^2 = 5.484$$

$\hat{\beta}_{(-i)}$ close to $\hat{\beta}$ unless the ith observation is an outlier. A $(1 - \alpha) \times 100\%$ confidence ellipsoid for β based on $\hat{\beta}$ is given by the set of all β^* such that

$$\frac{(\beta^* - \hat{\beta})\mathbf{X}'\mathbf{X}(\beta^* - \hat{\beta})}{(p + 1)\, MS_{\text{res}}} \leq F(1 - \alpha; p + 1, N - p - 1).$$

This ellipsoid is centered at $\hat{\beta}$.

Cook and Weisberg (1982, p. 118) indicate that a D_i of about 1, corresponding to distances between $\hat{\beta}$ and $\hat{\beta}_{(-i)}$ beyond a 50% confidence region, would generally be considered large. Cook's distance can be written in an alternative revealing form.

$$CD_i = \frac{1}{(p + 1)} r_i^2 \frac{h_{ii}}{1 - h_{ii}}, \qquad (11)$$

where r_i is the standardized residual and h_{ii} is the hat element. Thus, Cook's distance measures the joint (combined) influence on the case being an outlier on y and in the space of the predictors. A case may be influential because it is a significant outlier only on y, for example,

$$(p = 5, n = 40, r_i = 4, h_{ii} = .3 \Rightarrow CD_i > 1.),$$

or because it is a significant outlier only in the space of the predictors, for example,

$$(p = 5, n = 40, r_i = 2, h_{ii} = .7 \Rightarrow CD_i > 1.).$$

Note, however, that a case may not be a significant outlier on either y or in the space of the predictors, but may still be influential as in the following:

$$(p = 3, n = 20, h_{ii} = .4, r_i = 2.5 \Rightarrow CD_i > 1).$$

Example 1. Let us consider an example using data from Mickey, Dunn, and Clark (1967) which have $n = 19$ and one predictor. Points 1 and 3 are of particular interest from a diagnostic viewpoint. A variation is added on Point 1 (i.e., [42, 57]), changing it to (42, 30) for illustrative purposes. This is Point 2 in Figure 1.

When the regression analysis is run on their original data (labeled "Line without Point 2"), Point 3 is clearly an outlier on y ($r_i = 2.82$). However, it is not an influential data point because $h_{ii} = .057$ and $CD = .239$. Data Point 1 is an outlier on x because $D^2 = 10.604$. This D^2 exceeds the .05 critical value (8.73 for $p = 2$, $n = 18$) in Table 3. However, Cook's distance does not declare Point 1 as influential because $CD = .508 < 1$. The reason for this is that Cook's distance is influenced by the point's being an outlier on both y and on the predictor, and Point 1 is definitely not an outlier on y ($r = -.753$).

As discussed previously, the regression equation went from being significant to not significant when Point 1 was deleted. Thus, this point is certainly influential, but Cook's distance indicates otherwise. Hence, the rule that $CD_i > 1$ for a point to be influential should be used with some caution. As Cook and Weisberg (1982) note, "the ith case will be called influential if CD_i is large; the exact definition of large will depend on the problem, but $CD_i > 1$ . . . usually provides a basis for comparison" (p. 118).

All three diagnostics detect the point (42, 30) as influential. Cook's distance for this point (3.583) is much larger than it was for the point (42, 57). This is because the point is an outlier on both x and y ($r = 2.0$).

Finally, observe that the regression coefficients change substantially when Point 2 is deleted, from (118.64, −1.76) to (107.92, −.89). Because the Cook distance for Point 2 is very large (3.583), this is precisely what is expected, for CD_i measures how much the coefficients will change when the point is deleted.

Example 2. In the previous example a case was discussed where, when a single point is deleted, the regression coefficients may change considerably. The reader should be aware that there are also situations in which a group of cases is influential en bloc, but is not influential if considered individually. Cook and Weisberg (1980) have given a block of 2 points (C and D) that operate in this way (see Figure 3). If Point C or D is deleted, the regression line will change very little. However, if both are deleted, the fitted regression line will be quite different. Now consider the other block of points, A and B. If A or B is deleted, the fitted line will change; if both are deleted, the line will stay about the same.

Table 3
Critical Values for an Outlier in a Multivariate Normal Sample as Judged by Mahalanobis D^2

	$p = 2$		$p = 3$		$p = 4$		$p = 5$	
n	5%	1%	5%	1%	5%	1%	5%	1%
5	3.17	3.19						
6	4.00	4.11	4.14	4.16				
7	4.71	4.95	5.01	5.10	5.12	5.14		
8	5.32	5.70	5.77	5.97	6.01	6.09	6.11	6.12
9	5.85	6.37	6.43	6.76	6.80	6.97	7.01	7.08
10	6.32	6.97	7.01	7.47	7.50	7.79	7.82	7.98
12	7.10	8.00	7.99	8.70	8.67	9.20	9.19	9.57
14	7.74	8.84	8.78	9.71	9.61	10.37	10.29	10.90
16	8.27	9.54	9.44	10.56	10.39	11.36	11.20	12.02
18	8.73	10.15	10.00	11.28	11.06	12.20	11.96	12.98
20	9.13	10.67	10.49	11.91	11.63	12.93	12.62	13.81
25	9.94	11.73	11.48	13.18	12.78	14.40	13.94	15.47
30	10.58	12.54	12.24	14.14	13.67	15.51	14.95	16.73
35	11.10	13.20	12.85	14.92	14.37	16.40	15.75	17.73
40	11.53	13.74	13.36	15.56	14.96	17.13	16.41	18.55
45	11.90	14.20	13.80	16.10	15.46	17.74	16.97	19.24
50	12.23	14.60	14.18	16.56	15.89	18.27	17.45	19.83
100	14.22	16.95	16.45	19.26	18.43	21.30	20.26	23.17
200	15.99	18.94	18.42	21.47	20.59	23.72	22.59	25.82
500	18.12	21.22	20.75	23.95	23.06	26.37	25.21	28.62

n = number of observations; p = dimension.
Note. From *Outliers in Statistical Data* by V. Barnett and T. Lewis, 1978, New York: Wiley. Copyright 1978 by Wiley. Reprinted by permission of John Wiley & Sons, Ltd.

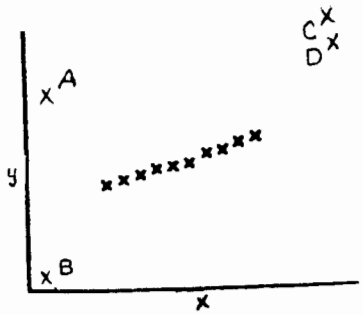


Figure 3. A group of cases which is influential en bloc, but not influential if considered individually. From "Characterizations of an empirical influence function for detecting influential cases in regression" by R. D. Cook and S. Weisberg, 1980, *Technometrics*, 22, p. 498. Copyright 1980 by American Statistical Association. Reprinted by permission.

Discussion

Because the results of a regression analysis may be seriously affected by just 1 or 2 errant data points, it is crucial for the researcher to isolate such points. The first source of errant points is recording errors, which can be detected by listing the data. Use of the Weisberg test (with studentized residuals) will detect y outliers, and the hat elements or the Mahalanobis distances will detect outliers in the space of the predictors. Such outliers will not necessarily be influential points. To determine which outliers are influential, find those whose Cook distances are >1. Those points that are flagged as influential by Cook's distance need to be examined carefully to determine whether they should be deleted from the analysis. If there is a reason to believe that these cases arise from a process different from that for the rest of the data, then the cases should be deleted. For example, the failure of a measuring instrument, a power failure, or the occurrence of an unusual event (perhaps inexplicable) would be instances of a different process.

If none of these appears to be the case, two analyses—one with the influential cases in and

one with these cases deleted—could be reported to emphasize the impact of these few points on the analysis.

References

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

Anscombe, F. J. (1973). Graphs in statistical analysis. *American Statistician, 27,* 17–21.

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data.* New York: Wiley.

Barr, A. J., Goodnight, J. H., Sall, J. P., & Heboig, J. T. (1982). *SAS user's guide: Statistics.* Raleigh, NC: Statistical Analysis System Institute.

Beckman, R. J., & Trussel, H. J. (1974). The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. *Journal of the American Statistical Association, 69,* 199–201.

Belsley, D. A., Kuh, E., & Welsch, R. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity.* New York: Wiley.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

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

Cook, R. D., & Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics, 22,* 495–508.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression.* New York: Chapman & Hall.

Dixon, W. J. (1981). *BMDP Statistical Software.* Berkeley, CA: University of California Press.

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

Draper, N. R., & Smith, H. (1981). *Applied regression analysis.* New York: Wiley.

Hoaglin, D., & Welsch, R. (1978). The hat matrix in regression and ANOVA. *American Statistician,* 17–22.

Hogg, R. V. (1979). Statistical robustness: One view of use in applications today. *American Statistician,* 108–115.

Huber, P. (1977). *Robust statistical procedures* (No. Regional conference series in applied mathem Philadelphia, PA: SIAM.

Hull, C. H., & Nie, N. H. (1981). *SPSS Update 7-9* York: McGraw-Hill.

Johnson, R., & Wichern, D. (1982). *Applied multiv statistical analysis.* Englewood Cliffs, NJ: Prentice-H

Johnson, W., & Geisser, S. (1983). A predictive vie the detection and characterization of influential ob vations in regression analysis. *Journal of the Amer Statistical Association, 78,* 137–144.

Krasker, W. S., & R. E. Welsch. (1979). *Efficient Bo Influence Regression Estimation Using Alternative initions of Sensitivity* (Tech. Rep. No. 3). Cambr MA: Center for Computational Research in Econo and Management Science, Massachusetts Institut Technology.

Mickey, M. R., Dunn, O. J., & Clark, V. (1967). Note the use of stepwise regression in detecting out *Computers and Biomedical Research, 1,* 105–111.

Mosteller, F., & Tukey, J. W. (1977). *Data analysis regression.* Reading, MA: Addison-Wesley.

Pedhazur, E. (1982). *Multiple regression in behavioral search.* New York: Holt, Rinehart & Winston.

Thissen, D., Baker, L., & Wainer, H. (1981). Influe enhanced scatter plots. *Psychological Bulletin, 90,* 184.

Tukey, J. W. (1977). *Exploratory data analysis.* Re MA: Addison-Wesley.

Weisberg, S. (1980). *Applied linear regression.* New Wiley.

Wilks, S. (1963). Multivariate statistical outliers. *Sank 25,* 507–526.

Received June 1, 19
Revision received August 18, 19

al Bulletin
l. 95, No. 2, 345-347

Data Trans
A Rebu

Division of

The Levine and Dunlap (difficult questions that mu transforming data to elim by using curvilinear transf that with large *n*s, the e transformation that will a unlikely. Thus, even if Le be applied to empirical d failure to distinguish theo about implications from unbiased test.

In their study, Levine and D made some very strong claims re and power. "Any skew reduces p results are quite general. . . . Th any skew in the data results in ratio, and the greater the lack the smaller the *F*, thus the grea power." (p. 279). This conclusio an empirical study using the th normal distribution, in which tre were added to lognormal distr log transformation was then ap treatment distributions. Natural the log transformation changes distributions to a normal curve the Levine and Dunlap generati the lognormal treatment popul treatment variances (p. 274) such transformation also was a varia transformation.

First, it should be noted that whether or not to transform important only for small sample sample sizes, the central limit notices that the sample means, distributed, and theoretical studies show that the analysis o 0%) and/or *t* tests work very

Requests for reprints should be sen Division of Counseling and Educatio 268 Rackley Building, Pennsylvania State U Park, Pennsylvania 16802.