Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
David A. Belsley, Edwin Kuh, Roy E. Welsch

ISBN: 0-471-05856-4

CHAPTER 2

Detecting Influential Observations and Outliers

In this chapter we identify subsets of the data that appear to have a disproportionate influence on the estimated model and ascertain which parts of the estimated model are most affected by these subsets. The focus is on methods that involve both the response (dependent) and the explanatory (independent) variables, since techniques not using both of these can fail to detect multivariate influential observations.

The sources of influential subsets are diverse. First, there is the inevitable occurrence of improperly recorded data, either at their source or in their transcription to computer-readable form. Second, observational errors are often inherent in the data. Although procedures more appropriate for estimation than ordinary least squares exist for this situation, the diagnostics we propose below may reveal the unsuspected existence and severity of observational errors. Third, outlying data points may be legitimately occurring extreme observations. Such data often contain valuable information that improves estimation efficiency by its presence. Even in this beneficial situation, however, it is constructive to isolate extreme points and to determine the extent to which the parameter estimates depend on these desirable data. Fourth, since the data could have been generated by a model(s) other than that specified, diagnostics may reveal patterns suggestive of these alternatives.

The fact that a small subset of the data can have a disproportionate influence on the estimated parameters or predictions is of concern to users of regression analysis, for, if this is the case, it is quite possible that the model-estimates are based primarily on this data subset rather than on the majority of the data. If, for example, the task at hand is the estimation of the mean and standard deviation of a univariate distribution, exploration

of the data will often reveal outliers, skewness, or multimodal distributions. Any one of these might cast suspicion on the data or the appropriateness of the mean and standard deviation as measures of location and variability, respectively. The original model may also be questioned, and transformations of the data consistent with an alternative model may be suggested. 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 cannot detect multivariate discrepant observations, nor can they tell us in what ways such data influence the estimated model.

After the multiple regression has been performed, most detection procedures focus on the residuals, the fitted (predicted) values, and the explanatory variables. Although much can be learned through such methods, they nevertheless fail to show us directly what the estimated model would be if a subset of the data were modified or set aside. Even if we are able to detect suspicious observations by these methods, we still will not know the extent to which their presence has affected the estimated coefficients, standard errors, and test statistics. In this chapter we develop techniques for diagnosing influential data points that avoid some of these weaknesses. In Section 2.1 the theoretical development is undertaken. Here new techniques are developed and traditional procedures are suitably modified and reinterpreted. In Section 2.2 the diagnostic procedures are exemplified through their use on an intercountry life-cycle savings function employing cross-sectional data. Further examples of these techniques and their interrelation with the collinearity diagnostics that are the subject of the next chapter are found in Chapter 4.

Before describing multivariate diagnostics, we present a brief two-dimensional graphic preview that indicates what sort of interesting situations might be subject to detection. We begin with an examination of Exhibit 2.1*a* which portrays a case that we might call (to avoid statistical connotations) evenly distributed. If the variance of the explanatory variable is small, slope estimates will often be unreliable, but in these circumstances standard test statistics contain the necessary information.

In Exhibit 2.1*b*, the point ∘ is anomalous, but since it occurs near the mean of the explanatory variable, no adverse effects are inflicted on the slope estimate. The intercept estimate, however, will be affected. The source of this discrepant observation might be in the response variable, or the error term. If it is the last, it could be indicative of heteroscedasticity or thick-tailed error distributions. Clearly, more such points are needed to analyze those problems fully, but isolating the single point is instructive.

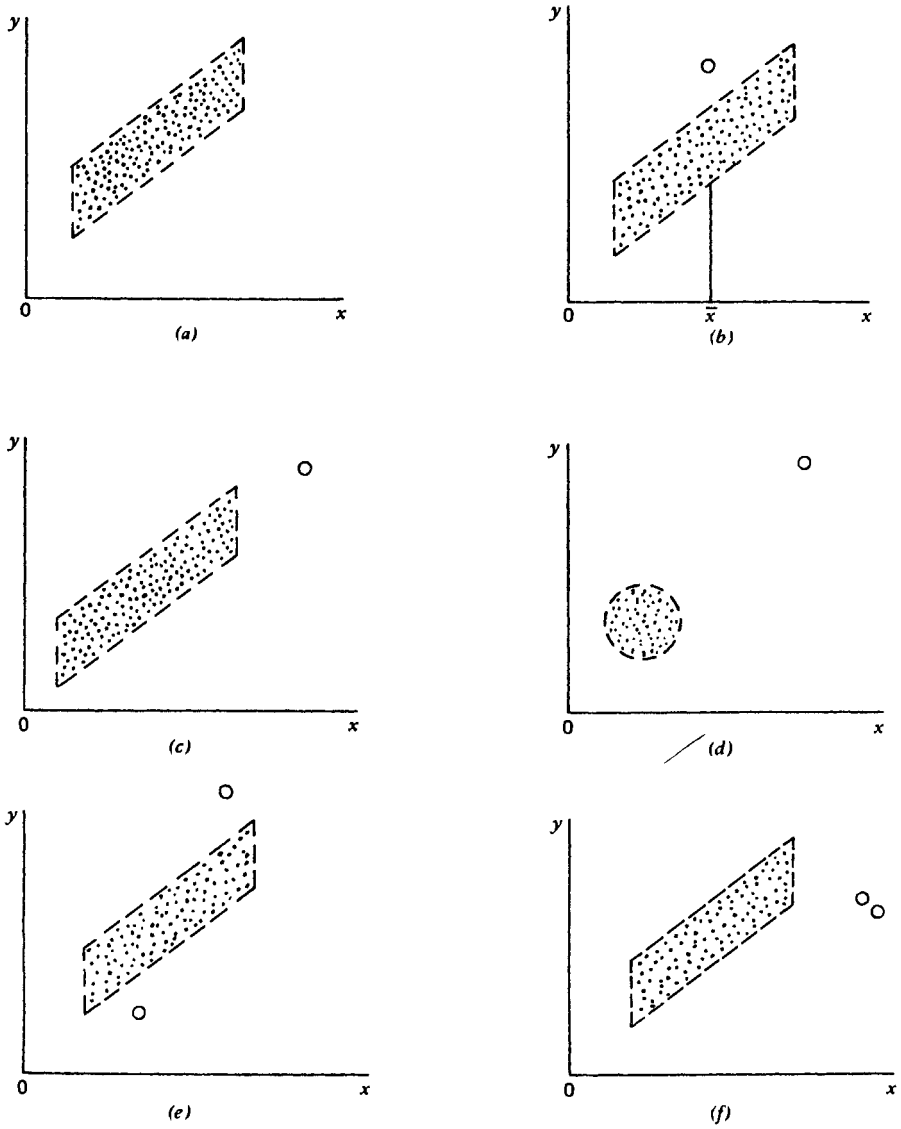

Exhibit 2.1 Plots for alternative configurations of data.

Exhibit 2.1*c* illustrates the case in which a gap separates the discrepant point from the main body of data. Since this potential outlier is consistent with the slope information contained in the rest of the data, this situation may exemplify the benevolent third source of influence mentioned above in which the outlying point supplies crucially useful information—in this case, a reduction in variance. Exhibit 2.1*d* is a more troublesome configuration that can arise frequently in practice. In this situation, the estimated regression slope is almost wholly determined by the extreme point. In its absence, the slope might be almost anything. Unless the extreme point is a crucial and valid piece of evidence (which, of course, depends on the research context), the researcher is likely to be highly suspicious of the estimate. Given the gap and configuration of the main body of data, the estimated slope surely has fewer than the usual degrees of freedom; in fact, it might appear that there are effectively only two data points.

The situation displayed in Exhibit 2.1*e* is a potential source of concern since either or both ∘'s will heavily influence the slope estimate, but differently from the remaining data. Here is a case where some corrective action is clearly indicated—either data deletion or, less drastically, a downweighting of the suspicious observations or possibly even a model reformulation.

Finally, Exhibit 2.1*f* presents an interesting case in which deletion of either ∘ by itself will have little effect on the regression outcome. The potential effect of one outlying observation is clearly being masked by the presence of the other. This example serves as simple evidence for the need to examine the effects of more general subsets of the data.

2.1 THEORETICAL FOUNDATIONS

In this section we present the technical background for diagnosing influential data points. Our discussion begins with a description of the technique of row deletion, at first limited to deleting one row (observation) at a time. This procedure is easy to understand and to compute. Here we examine in turn how the deletion of a single row affects the estimated coefficients, the predicted (fitted) values, the residuals, and the estimated covariance structure of the coefficients. These four outputs of the estimation process are, of course, most familiar to users of multiple regression and provide a basic core of diagnostic tools.

The second diagnostic procedure is based on derivatives of various regression outputs with respect to selected regression inputs. In particular, it proves useful to examine the sensitivity of the regression output to small

perturbations away from the usual regression assumption of homoscedasticity. Elements of the theory of robust estimation can then be used to convert these derivatives into diagnostic measures.

The third diagnostic technique moves away from the traditional regression framework and focuses on a geometric approach. The y vector is adjoined to the X matrix to form n data points in a $p+1$ dimensional space. It then becomes possible for multivariate methods, such as ratios of determinants, to be used to diagnose discrepant points. The emphasis here is on locating outliers in a geometric sense.

Our attention then turns to more comprehensive diagnostic techniques that involve the deletion or perturbation of more than one row at a time. Such added complications prove necessary, for, in removing only one row at a time, the influence of a group of influential observations may not be adequately revealed. Similarly, an influential data point that coexists with others may have its influence masked by their presence, and thus remain hidden from detection by single-point (one-at-a-time) diagnostic techniques. The first multiple-point (more-than-one-at-a-time) procedures we examine involve the deletion of subsets of data, with particular emphasis on the resulting change in coefficients and fitted values. Since multiple deletion is relatively expensive, lower-cost stepwise[1] methods are also introduced.

The next class of procedures adjoins to the X matrix a set of dummy variables, one for each row under consideration. Each dummy variate consists of all zeros except for a one in the appropriate row position. Variable-selection techniques, such as stepwise regression or regressions with all possible subsets removed, can be used to select the discrepant rows by noting which dummy variables remain in the regression. The derivative approaches can also be generalized to multiple rows. The emphasis is placed both on procedures that perturb the homoscedasticity assumption in exactly the same way for all rows in a subset and on low-cost stepwise methods.

Next we examine the usefulness of Wilks' Λ statistic applied to the matrix Z, formed by adjoining y to X, as a means for diagnosing groups of outlying observations. This turns out to be especially useful either when there is no natural way to form groups, as with most cross-sectional data, or when unexpected groupings occur, such as might be the case in census tract data. We also examine the Andrews-Pregibon (1978) statistic.

[1]The use of the term *stepwise* in this context should not be confused with the concept of stepwise regression, which is not being indicated. The term *sequential* was considered but not adopted because of its established statistical connotations.

Finally we consider generalized distance measures (like the Mahalanobis distance) applied to the **Z** matrix. These distances are computed in a stepwise manner, thus allowing more than one row at a time to be considered.

A useful summary of the notation employed is given in Exhibit 2.2.

Single-Row Effects

We develop techniques here for discovering influential observations.[2] Each observation, of course, is closely associated with a single row of the data matrix **X** and the corresponding element of **y**.[3] An influential observation is one which, either individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates (coefficients, standard errors, t-values, etc.) than is the case for most of the other observations. One obvious means for examining such an impact is to delete each row, one at a time, and note the resultant effect on the various calculated values.[4] Rows whose deletion produces relatively large changes in the calculated values are deemed influential. We begin, then, with an examination of this procedure of row deletion, looking in turn at the impact of each row on the estimated coefficients and the predicted (fitted) values ($\hat{y}$'s), the residuals, and the estimated parameter variance-covariance matrix. We then turn to other means of locating single data points with high impact: differentiation of the various calculated values with respect to the weight attached to an observation, and a geometrical view based on distance measures. Generalizations of some of these procedures to the problem of assessing the impact of deleting more than one row at a time are then examined.

Deletion.

Coefficients and Fitted Values. Since the estimated coefficients are often of primary interest to users of regression models, we look first at the change in the estimated regression coefficients that would occur if the ith row were deleted. Denoting the coefficients estimated with the ith row

[2]A number of the concepts employed in this section have been drawn from the existing literature. Relevant citations accompany the derivation of these formulae in Appendix 2A.

[3]Observations and rows need not be uniquely paired, for in time-series models with lagged variables, the data relevant to a given observation could occur in several neighboring rows. We defer further discussion of this aspect of time-series data until Chapters 4 and 5, and continue here to use these two terms interchangeably.

[4]The term *row deletion* is used generally to indicate the deletion of a row from both the **X** matrix and the **y** vector.

Exhibit 2.2 Notational conventions

Population Regression $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$		Estimated Regression $\mathbf{y}=\mathbf{X}\mathbf{b}+\mathbf{e}$	
$\mathbf{y}$:	$n\times 1$ column vector for response variable	same	
$\mathbf{X}$:	$n\times p$ matrix of explanatory variables*	same	
$\boldsymbol{\beta}$:	$p\times 1$ column vector of regression parameters	$\mathbf{b}$:	estimate of $\boldsymbol{\beta}$
$\boldsymbol{\epsilon}$:	$n\times 1$ column vector of errors	$\mathbf{e}$:	residual vector
σ^2:	error variance	s^2:	estimated error variance
	Additional Notation		
$\mathbf{x}_i$:	ith row of $\mathbf{X}$ matrix	$\mathbf{b}(i)$:	estimate of $\boldsymbol{\beta}$ when ith row of $\mathbf{X}$ and $\mathbf{y}$ have been deleted.
$\mathbf{X}_j$:	jth column of $\mathbf{X}$ matrix	$s^2(i)$:	estimated error variance when ith row of $\mathbf{X}$ and $\mathbf{y}$ have been deleted.
$\mathbf{X}(i)$:	$\mathbf{X}$ matrix with ith row deleted.		

Matrices are transposed with a superscript T, as in $\mathbf{X}^T\mathbf{X}$. Mention should also be made of a convention that is adopted in the reporting of regression results. Estimated standard errors of the regression coefficients are always reported in parentheses beneath the corresponding coefficient. In those cases where emphasis is on specific tests of significance, the t's are reported instead, and are always placed in square brackets. Other notation is either obvious or is introduced in its specific context.

*We typically assume $\mathbf{X}$ to contain a column of ones, corresponding to the constant term. In the event that $\mathbf{X}$ contains no such column, certain of the formulas must have their degrees of freedom altered accordingly. In particular, at a latter stage we introduce the notation $\tilde{\mathbf{X}}$ to indicate the matrix formed by centering the columns of $\mathbf{X}$ about their respective column means. If the $n \times p$ matrix $\mathbf{X}$ contains a constant column of ones, $\tilde{\mathbf{X}}$ is assumed to be of size $n\times(p-1)$, the column of zeros being removed. The formulas as written take into account this change in degrees of freedom. Should $\mathbf{X}$ contain no constant column, however, all formulas dealing with centered matrices must have their degrees of freedom increased by one.

deleted by $\mathbf{b}(i)$, this change is easily computed from the formula

$$\text{DFBETA}_i \equiv \mathbf{b} - \mathbf{b}(i) = \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T e_i}{1-h_i}, \tag{2.1}$$

where

$$h_i = \mathbf{x}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T, \tag{2.2}$$

and the reader is reminded that $\mathbf{x}_i$ is a *row* vector. The quantity h_i occurs frequently in the diagnostics developed in this chapter and it is discussed more below.[5]

Whether the change in b_j, the jth component of $\mathbf{b}$, that results from the deletion of the ith row is large or small is often most usefully assessed relative to the variance of b_j, that is, $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}$. If we let

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \tag{2.3}$$

then

$$b_j - b_j(i) = \frac{c_{ji}e_i}{1-h_i}. \tag{2.4}$$

Since

$$\sum_{i=1}^{n} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T\mathbf{x}_i(\mathbf{X}^T\mathbf{X})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}, \tag{2.5}$$

it follows that [see Mosteller and Tukey (1977)]

$$\text{var}(b_j) = \sigma^2 \sum_{k=1}^{n} c_{jk}^2. \tag{2.6}$$

Thus a scaled measure of change can be defined as

$$\text{DFBETAS}_{ij} \equiv \frac{b_j - b_j(i)}{s(i)\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} = \frac{c_{ji}}{\sqrt{\sum_{k=1}^{n} c_{jk}^2}} \frac{e_i}{s(i)(1-h_i)}, \tag{2.7}$$

[5]See Appendixes 2A and 2B for details on the computation of the h_i.

where we have replaced s^2, the usual estimate of σ^2, by

$$s^2(i)=\frac{1}{n-p-1}\sum_{k\neq i}\left[y_k-\mathbf{x}_k\mathbf{b}(i)\right]^2$$

in order to make the denominator stochastically independent of the numerator in the Gaussian (normal) case. A simple formula for $s(i)$ results from

$$(n-p-1)s^2(i)=(n-p)s^2-\frac{e_i^2}{1-h_i}. \tag{2.8}$$

In the special case of location,

$$\text{DFBETA}_i=\frac{e_i}{n-1}$$

and

$$\text{DFBETAS}_i=\frac{\sqrt{n}\,e_i}{(n-1)s(i)}. \tag{2.9}$$

As we might expect, the chance of getting a large DFBETA is reduced in direct proportion to the increase in sample size. Deleting one observation should have less effect as the sample size grows. Even though scaled by a measure of the standard error of b, DFBETAS_i decreases in proportion to $\sqrt{n}$.

Returning to the general case, large values of $|\text{DFBETAS}_{ij}|$ indicate observations that are influential in the determination of the jth coefficient, b_j.[6] The nature of "large" in relation to the sample size, n, is discussed below.

Another way to summarize coefficient changes and, at the same time, to gain insight into forecasting effects when an observation is deleted is by

[6] When the Gaussian assumption holds, it can also be useful to look at the change in t-statistics as a means for assessing the sensitivity of the regression output to the deletion of the ith row, that is, to examine

$$\text{DFTSTAT}_{ij}\equiv\frac{b_j}{s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}}-\frac{b_j(i)}{s(i)\sqrt{[\mathbf{X}^T(i)\mathbf{X}(i)]^{-1}_{jj}}}.$$

Studying the changes in regression statistics is a good second-order diagnostic tool because, if a row appears to be overly influential on other grounds, an examination of the regression statistics will show whether the conclusions of hypothesis testing would be affected.

the change in fit, defined as

$$\text{DFFIT}_i \equiv \hat{y}_i - \hat{y}_i(i) = \mathbf{x}_i[\mathbf{b} - \mathbf{b}(i)] = \frac{h_i e_i}{1-h_i}. \tag{2.10}$$

This diagnostic measure has the advantage that it does not depend on the particular coordinate system used to form the regression model. For scaling purposes, it is natural to divide by $\sigma\sqrt{h_i}$, the standard deviation of the fit, $\hat{y}_i = \mathbf{x}_i\mathbf{b}$, giving

$$\text{DFFITS}_i \equiv \left[\frac{h_i}{1-h_i}\right]^{1/2} \frac{e_i}{s(i)\sqrt{1-h_i}}, \tag{2.11}$$

where σ has been estimated by $s(i)$. A measure similar to (2.11) has been suggested by Cook (1977).

It is natural to ask about the scaled changes in fit for other than the ith row; that is,

$$\frac{\mathbf{x}_k(\mathbf{b}-\mathbf{b}(i))}{s(i)\sqrt{h_k}} = \frac{h_{ik}e_i}{s(i)\sqrt{h_k}\,(1-h_i)}, \tag{2.12}$$

where $h_{ik} \equiv \mathbf{x}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_k^T$. Since

$$\sup_{\boldsymbol{\lambda}} \frac{|\boldsymbol{\lambda}^T[\mathbf{b}-\mathbf{b}(i)]|}{s(i)\left[\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\lambda}\right]^{1/2}} = \frac{\left\{[\mathbf{b}-\mathbf{b}(i)]^T(\mathbf{X}^T\mathbf{X})[\mathbf{b}-\mathbf{b}(i)]\right\}^{1/2}}{s(i)}$$

$$\equiv |\text{DFFITS}_i|, \tag{2.13}$$

it follows that

$$\left|\frac{\mathbf{x}_k[\mathbf{b}-\mathbf{b}(i)]}{s(i)\sqrt{h_k}}\right| \leq |\text{DFFITS}_i|. \tag{2.14}$$

Thus $|\text{DFFITS}_i|$ dominates the expression in (2.12) for all k and these latter measures need only be investigated when $|\text{DFFITS}_i|$ is large.

A word of warning is in order here, for it is obvious that there is room for misuse of the above procedures. High-influence data points could conceivably be removed solely to effect a desired change in a particular estimated coefficient, its t-value, or some other regression output. While

this danger surely exists, it is an unavoidable consequence of a procedure that successfully highlights such points. It should be obvious that an influential point is legitimately deleted altogether only if, once identified, it can be shown to be uncorrectably in error. Often no action is warranted, and when it is, the appropriate action is usually more subtle than simple deletion. Examples of corrective action are given in Section 2.2 and in Chapter 4. These examples show that the benefits obtained from information on influential points far outweigh any potential danger.

The Hat Matrix. Returning now to our discussion of deletion diagnostics, we can see from (2.1) to (2.11) that h_i and e_i are fundamental components. Some special properties of h_i are discussed in the remainder of this section and we study special types of residuals (like $e_i/s(i)\sqrt{1-h_i}$) in the next section.[7]

The h_i are the diagonal elements of the least-squares projection matrix, also called the hat matrix,

$$\mathbf{H}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \tag{2.15}$$

which determines the fitted or predicted values, since

$$\hat{\mathbf{y}}\equiv\mathbf{X}\mathbf{b}=\mathbf{H}\mathbf{y}. \tag{2.16}$$

The influence of the response value, y_i, on the fit is most directly reflected in its impact on the corresponding fitted value, $\hat{y}_i$, and this information is seen from (2.16) to be contained in h_i. The diagonal elements of $\mathbf{H}$ can also be related to the distance between $\mathbf{x}_i$ and $\bar{\mathbf{x}}$ (the row vector of explanatory variable means). Denoting by tilde data that have been centered, we show in Appendix 2A that

$$h_i-\frac{1}{n}=\tilde{h}_i=\tilde{\mathbf{x}}_i(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{x}}_i^T. \tag{2.17}$$

We see from (2.17) that $\tilde{h}_i$ is a positive-definite quadratic form and thus possesses an appropriate distance interpretation.[8]

Where there are two or fewer explanatory variables, scatter plots will quickly reveal any x-outliers, and it is not hard to verify that they have

[7]The immediately following material closely follows Hoaglin and Welsch (1978).

[8]As is well known [Rao (1973), Section 1c.1], any $n\times n$ positive-definite matrix $\mathbf{A}$ may be decomposed as $\mathbf{A}=\mathbf{P}^T\mathbf{P}$ for some non-singular matrix $\mathbf{P}$. Hence the positive-definite quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$ ($\mathbf{x}$ an n-vector) is equivalent to the sum of squares $\mathbf{z}^T\mathbf{z}$ (the squared Euclidean length of the n-vector $\mathbf{z}$), where $\mathbf{z}=\mathbf{P}\mathbf{x}$.

relatively large h_i values. When $p>2$, scatter plots may not reveal "multivariate outliers," which are separated from the bulk of the x-points but do not appear as outliers in a plot of any single explanatory variable or pair of them. Since, as we have seen, the diagonal elements of the hat matrix **H** have a distance interpretation, they provide a basic starting point for revealing such "multivariate outliers." These diagonals of the hat matrix, the h_i, are diagnostic tools in their own right as well as being fundamental parts of other such tools.

H is a projection matrix and hence, as is shown in Appendix 2A,

$$0 \leqslant h_i \leqslant 1. \tag{2.18}$$

Further, since **X** is of full rank,

$$\sum_{i=1}^{n} h_i = p. \tag{2.19}$$

The average size of a diagonal element, then, is p/n. Now if we were designing an experiment, it would be desirable to use data that were roughly equally influential, that is, each observation having an h_i near to the average p/n. But since the **X** data are usually given to us, we need some criterion to decide when a value of h_i is large enough (far enough away from the average) to warrant attention.

When the explanatory variables are independently distributed as the multivariate Gaussian, it is possible to compute the exact distribution of certain functions of the h_i's. We use these results for guidance only, realizing that independence and the Gaussian assumption are often not valid in practice. In Appendix 2A, $(n-p)[h_i-(1/n)]/(1-h_i)(p-1)$ is shown to be distributed as F with $p-1$ and $n-p$ degrees of freedom. For $p>10$ and $n-p>50$ the 95% value for F is less than 2 and hence $2p/n$ (twice the balanced average h_i) is a good rough cutoff. When $p/n>0.4$, there are so few degrees of freedom per parameter that all observations become suspect. For small p, $2p/n$ tends to call a few too many points to our attention, but it is simple to remember and easy to use. In what follows, then, we call the ith observation a *leverage point* when h_i exceeds $2p/n$. The term *leverage* is reserved for use in this context.

Note that when $h_i=1$, we have $\hat{y}_i=y_i$; that is, $e_i=0$. This is equivalent to saying that, in some coordinate system, one parameter is determined completely by y_i or, in effect, dedicated to one data point. A proof of this result is given in Appendix 2A where it is also shown that

$$\det\left[\mathbf{X}^T(i)\mathbf{X}(i)\right] = (1-h_i)\det(\mathbf{X}^T\mathbf{X}). \tag{2.20}$$

Clearly when $h_i = 1$ the new matrix $\mathbf{X}(i)$, formed by deleting the ith row, is singular and we cannot obtain the usual least-squares estimates. This is extreme leverage and does not often occur in practice.

We complete our discussion of the hat matrix with a few simple examples. For the sample mean, all elements of H are $1/n$. Here $p = 1$ and each $h_i = p/n$, the perfectly balanced case.

For a straight line through the origin,

$$h_{ij} = \frac{x_i x_j}{\sum_{k=1}^{n} x_k^2}, \tag{2.21}$$

and

$$\sum_{i=1}^{n} h_i = p = 1.$$

Simple linear regression is slightly more complicated, but a few steps of algebra give

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})^2} \tag{2.22}$$

Residuals. We turn now to an examination of the diagnostic value of the effects that deleting rows can have on the regression residuals. The use of the regression residuals in a diagnostic context is, of course, not new. Looking at regression residuals, $e_i = y_i - \hat{y}_i$, and especially large residuals, has traditionally been used to highlight data points suspected of unduly affecting regression results. The residuals have also been employed to detect departures from the Gauss-Markov assumptions on which the desirable properties of least squares rest. As is well known, the residuals can be used to detect some forms of heteroscedasticity and auto-correlation, and can provide the basis for mitigating these problems. The residuals can also be used to test for the approximate normality of the disturbance term. Since the least-squares estimates retain their property of best-linear-unbiasedness even in the absence of normality of the disturbances, such tests are often overlooked in econometric practice, but even moderate departures from normality can noticeably impair estimation efficiency[9] and the meaningfulness of standard tests of hypotheses. Harmful departures from normality include pronounced skewness, multiple modes, and thick-tailed distributions. In all these uses of residuals,

[9] The term efficiency is used here in a broad sense to indicate minimum mean-squared error.

one should bear in mind that large outliers among the true errors, ϵ_i, can often be reflected in only modest-sized least-squares residuals, since the squared-error criterion weights extreme values heavily.

Three diagnostic measures based on regression residuals are presented here; two deal directly with the estimated residuals and the third results from a change in the assumption on the error distribution. The first is simply a frequency distribution of the residuals. If there is evident visual skewness, multiple modes, or a heavy-tailed distribution, a graph of the frequency distribution will prove informative. It is worth noting that economists often look at time plots of residuals but seldom at their frequency or cumulative distribution.

The second is the normal probability plot, which displays the cumulative normal distribution as a straight line whose slope measures the standard deviation and whose intercept reflects the mean. Thus a failure of the residuals to be normally distributed will often reveal itself as a departure of the cumulative residual plot from a straight line. Outliers often appear at either end of the cumulative distribution.

Finally, Denby and Mallows (1977) and Welsch (1976) have suggested plotting the estimated coefficients and residuals as the error likelihood, or, equivalently, as the criterion function (negative logarithm of the likelihood) is changed. One such family of criterion functions has been suggested by Huber (1973); namely,

$$\rho_c(t)=\begin{cases} \dfrac{t^2}{2} & |t|\leqslant c \\ c|t|-\dfrac{c^2}{2} & |t|>c, \end{cases} \tag{2.23}$$

which goes from least squares ($c=\infty$) to least absolute residuals ($c\to 0$). This approach is attractive because of its relation to robust estimation, but it requires considerable computation.

For diagnostic use the residuals can be modified in ways that will enhance our ability to detect problem data. It is well known [Theil (1971)] that

$$\operatorname{var}(e_i)=\sigma^2(1-h_i). \tag{2.24}$$

Consequently, many authors have suggested that, instead of studying e_i, we should use the *standardized residuals*

$$e_{si}\equiv\frac{e_i}{s\sqrt{1-h_i}}. \tag{2.25}$$

We prefer instead to estimate σ by $s(i)$ [cf. (2.8)]. The result is a *studentized residual* (RSTUDENT),

$$e_i^* \equiv \frac{e_i}{s(i)\sqrt{1-h_i}}, \tag{2.26}$$

which, in a number of practical situations, is distributed closely to the t-distribution with $n-p-1$ degrees of freedom. Thus, if the Gaussian assumption holds, we can readily assess the significance of any single studentized residual. Of course, the e_i^* will not be independent.

The studentized residuals have another interesting interpretation. If we were to add to the data a dummy variable consisting of a column with all zeros except for a one in the ith row (the new model), then e_i^* is the t-statistic that tests for the significance of the coefficient of this new variable. To prove this, let SSR stand for sum of squared residuals and note that

$$\frac{[\text{SSR(old model)} - \text{SSR(new model)}]/1}{\text{SSR(new model)}/(n-p-1)} \tag{2.27}$$

$$= \frac{(n-p)s^2 - (n-p-1)s^2(i)}{s^2(i)} = \frac{e_i^2}{s^2(i)(1-h_i)}. \tag{2.28}$$

Under the Gaussian assumption, (2.27) is distributed as $F_{1,n-p-1}$, and the result follows by taking the square root of (2.28). Some additional details are contained in Appendix 2A.

The studentized residuals thus provide a better way to examine the information in the residuals, both because they have equal variances and because they are easily related to the t-distribution in many situations. However, this does not tell the whole story, since some of the most influential data points can have relatively small studentized residuals (and very small e_i).

To illustrate with the simplest case, regression through the origin, note that

$$b - b(i) = \frac{x_i e_i}{\sum_{j \neq i} x_j^2}. \tag{2.29}$$

Equation (2.29) shows that the residuals are related to the change in the least-squares estimate caused by deleting one row, but each contains different information, since large values of $|b - b(i)|$ can be associated with

small $|e_i|$ and vice versa. Hence row deletion and the analysis of residuals need to be treated together and on an equal footing.

When the index of observations is time, the studentized residuals can be related to the recursive residuals proposed by Brown, Durbin, and Evans (1975). If $\mathbf{b}(t)$ is the least-squares estimate based on the first $t-1$ observations, then the recursive residuals are defined to be

$$q_t = \frac{y_t - \mathbf{x}_t\mathbf{b}(t)}{\left\{1+\mathbf{x}_t\left[\mathbf{X}^T(t)\mathbf{X}(t)\right]^{-1}\mathbf{x}_t^T\right\}^{1/2}}, \qquad t=p+1,\ldots,T. \tag{2.30}$$

which by simple algebra (see Appendix 2A) can be written as

$$\frac{y_t - \mathbf{x}_t\mathbf{b}}{\sqrt{1-h_t}}, \tag{2.31}$$

where h_t and $\mathbf{b}$ are computed from the first t observations. For a related interpretation see a discussion of the PRESS residual by Allen (1971).

When we set

$$S_t \equiv \sum_{i=1}^{t} (y_i - \mathbf{x}_i\mathbf{b})^2, \tag{2.32}$$

(2.8) gives

$$S_t = S_{t-1} + q_t^2. \tag{2.33}$$

Brown, Durbin, and Evans propose two tests for studying the constancy of regression relationships over time. The first uses the cusum

$$W_t = \frac{T-p}{S_T}\sum_{j=p+1}^{t} q_j, \qquad t=p+1,\ldots,T, \tag{2.34}$$

and the second the cusum-of-squares

$$c_t = \frac{S_t}{S_T}, \qquad t=p+1,\ldots,T. \tag{2.35}$$

Schweder (1976) has shown that certain modifications of these tests, obtained by summing from $j=T$ to $t \geqslant p+1$ (backward cusum, etc.) have greater average power. The reader is referred to that paper for further details. An example of the use of these tests is given in Section 4.3.

Covariance Matrix. So far we have focused on coefficients, predicted (fitted) values of y, and residuals. Another major aspect of regression is the covariance matrix of the estimated coefficients.[10] We again consider the diagnostic technique of row deletion, this time in a comparison of the covariance matrix using all the data, $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, with the covariance matrix that results when the ith row has been deleted, $\sigma^2[\mathbf{X}^T(i)\mathbf{X}(i)]^{-1}$. Of the various alternative means for comparing two such positive-definite symmetric matrices, the ratio of their determinants $\det[\mathbf{X}^T(i)\mathbf{X}(i)]^{-1}/\det(\mathbf{X}^T\mathbf{X})^{-1}$ is one of the simplest and, in the present application, is quite appealing. Since these two matrices differ only by the inclusion of the ith row in the sum of squares and cross products, values of this ratio near unity can be taken to indicate that the two covariance matrices are close, or that the covariance matrix is insensitive to the deletion of row i. Of course, the preceding analysis is based on information from the X matrix alone and ignores the fact that the estimator s^2 of σ^2 also changes with the deletion of the ith observation. We can bring the y data into consideration by comparing the two matrices $s^2(\mathbf{X}^T\mathbf{X})^{-1}$ and $s^2(i)[\mathbf{X}^T(i)\mathbf{X}(i)]^{-1}$ in the determinantal ratio,

$$\begin{aligned}\text{COVRATIO} &\equiv \frac{\det\left\{s^2(i)\left[\mathbf{X}^T(i)\mathbf{X}(i)\right]^{-1}\right\}}{\det\left[s^2(\mathbf{X}^T\mathbf{X})^{-1}\right]} \\ &= \frac{s^{2p}(i)}{s^{2p}}\left\{\frac{\det\left[\mathbf{X}^T(i)\mathbf{X}(i)\right]^{-1}}{\det(\mathbf{X}^T\mathbf{X})^{-1}}\right\}. \qquad (2.36)\end{aligned}$$

Equation (2.36) may be given a more useful formulation by applying (2.20) to show

$$\frac{\det\left[\mathbf{X}^T(i)\mathbf{X}(i)\right]^{-1}}{\det(\mathbf{X}^T\mathbf{X})^{-1}} = \frac{1}{1-h_i}. \qquad (2.37)$$

Hence, using (2.8) and (2.26) we have

$$\text{COVRATIO} = \frac{1}{\left[\dfrac{n-p-1}{n-p} + \dfrac{e_i^{*2}}{n-p}\right]^p (1-h_i)}. \qquad (2.38)$$

[10]A diagnostic based on the diagonal elements of the covariance matrix can be obtained from the expression (2.6). By noting which c_{ji}^2 appear to be excessively large for a given j, we determine those observations that influence the variance of the jth coefficient. This diagnostic, however, has two weaknesses. First, it ignores the off-diagonal elements of the covariance matrix and second, emphasis on the c_{ji}^2 ignores s^2.

As a diagnostic tool, then, we are interested in observations that result in values of COVRATIO from (2.38) that are not near unity, for these observations are possibly influential and warrant further investigation.

In order to provide a rough guide to the magnitude of such variations from unity, we consider the two extreme cases $|e_i^*| > 2$ with h_i at its minimum $(1/n)$ and $h_i \geqslant 2p/n$ with $e_i^* = 0$. In the first case we get

$$\text{COVRATIO} \approx \frac{1}{\left(1+\dfrac{e_i^{*2}-1}{n-p}\right)^p} < \frac{1}{\left(1+\dfrac{3}{n-p}\right)^p}.$$

Further approximation leads to

$$\frac{1}{\left(1+\dfrac{3}{n-p}\right)^p} \approx \left(1+\frac{3p}{n}\right)^{-1} \approx 1-\frac{3p}{n}, \tag{2.39}$$

where $n-p$ has been replaced by n for simplicity. The latter bounds are, of course, not useful when $n \leqslant 3p$. For the second case

$$\text{COVRATIO} \approx \frac{1}{\left(1-\dfrac{1}{n-p}\right)^p} \frac{1}{(1-h_i)} > \frac{1}{\left(1-\dfrac{1}{n-p}\right)^p \left(1-\dfrac{2p}{n}\right)}.$$

A cruder but simpler bound follows from

$$\frac{1}{\left(1-\dfrac{1}{n-p}\right)^p \left(1-\dfrac{2p}{n}\right)} \approx \frac{1}{\left(1-\dfrac{p}{n-p}\right)\left(1-\dfrac{2p}{n}\right)}$$

$$\approx \left(1-\frac{3p}{n}\right)^{-1} \approx 1+\frac{3p}{n}. \tag{2.40}$$

Therefore we investigate points with $|\text{COVRATIO}-1|$ near to or larger than $3p/n$.[11]

The formula in (2.38) is a function of basic building blocks, such as h_i and the studentized residuals. Roughly speaking (2.38) will be large when h_i is large and small when the studentized residual is large. Clearly those

[11] Some prefer to normalize expressions like (2.36) for model size by raising them to the $1/p$th power. Had such normalization been done here, the approximations corresponding to (2.39) and (2.40) would be $1-(3/n)$ and $1+(3/n)$ respectively.

two factors can offset each other and that is why it is useful to look at them separately and in combinations as in (2.38).

We are also interested in how the variance of $\hat{y}_i$ changes when an observation is deleted. To do this we compute

$$\text{var}(\hat{y}_i) = s^2 h_i$$

$$\text{var}(\hat{y}_i(i)) = \text{var}(\mathbf{x}_i \mathbf{b}(i)) = s^2(i)\left[\frac{h_i}{1-h_i}\right],$$

and form the ratio

$$\text{FVARATIO} \equiv \frac{s^2(i)}{s^2(1-h_i)}.$$

This expression is similar to COVRATIO except that $s^2(i)/s^2$ is not raised to the pth power. As a diagnostic measure it will exhibit the same patterns of behavior with respect to different configurations of h_i and the studentized residual as described above for COVRATIO.

Differentiation. We examine now a second means for identifying influential observations, differentiation of regression outputs with respect to specific model parameters. In particular, we can alter the weight attached to the ith observation if, in the assumptions of the standard linear regression model, we replace $\text{var}(\epsilon_i) = \sigma^2$ with $\text{var}(\epsilon_i) = \sigma^2/w_i$, for the specific i only. Differentiation of the regression coefficients with respect to w_i, evaluated at $w_i = 1$, provides a means for examining the sensitivity of the regression coefficients to a slight change in the weight given to the ith observation. Large values of this derivative indicate observations that have large influence on the calculated coefficients. This derivative, as is shown in Appendix 2A, is

$$\frac{\partial \mathbf{b}(w_i)}{\partial w_i} = \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T e_i}{\left[1-(1-w_i)h_i\right]^2}, \tag{2.41}$$

and it follows that

$$\left.\frac{\partial \mathbf{b}(w_i)}{\partial w_i}\right|_{w_i=1} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T e_i. \tag{2.42}$$

This last formula is often viewed as the influence of the ith observation on

the estimated coefficients. Its relationship to the formula (2.1) for DFBETA is obvious and it could be used as an alternative to that statistic.

The theory of robust estimation [cf. Huber (1973)] implies that influence functions such as (2.42) can be used to approximate the covariance matrix of **b** by forming

$$\sum_{i=1}^{n} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T e_i e_i \mathbf{x}_i (\mathbf{X}^T\mathbf{X})^{-1} = \sum_{i=1}^{n} e_i^2 (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T\mathbf{x}_i (\mathbf{X}^T\mathbf{X})^{-1}. \quad (2.43)$$

This is not quite the usual covariance matrix, but if e_i^2 is replaced by the average value, $\sum_{k=1}^{n} e_k^2/n$, we get

$$\frac{\sum_{k=1}^{n} e_k^2}{n} \sum_{i=1}^{n} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T\mathbf{x}_i (\mathbf{X}^T\mathbf{X})^{-1} = \frac{\sum_{k=1}^{n} e_k^2}{n} (\mathbf{X}^T\mathbf{X})^{-1}, \quad (2.44)$$

which, except for degrees of freedom, is the estimated least-squares covariance matrix.

To assess the influence of an individual observation, we compare

$$\sum_{k \neq i} e_k^2 (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_k^T\mathbf{x}_k (\mathbf{X}^T\mathbf{X})^{-1} \quad (2.45)$$

with

$$s^2(\mathbf{X}^T\mathbf{X})^{-1}. \quad (2.46)$$

The use of determinants with the sums in (2.45) is difficult, so we replace e_k^2 for $k \neq i$ by $s^2(i)$, leaving

$$s^2(i) \sum_{k \neq i} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_k^T\mathbf{x}_k (\mathbf{X}^T\mathbf{X})^{-1} = s^2(i)(\mathbf{X}^T\mathbf{X})^{-1}\left[\mathbf{X}^T(i)\mathbf{X}(i)\right](\mathbf{X}^T\mathbf{X})^{-1}. \quad (2.47)$$

Forming the ratio of the determinant of (2.47) to that of (2.46) we get

$$\frac{s^{2p}(i)}{s^{2p}} \cdot \frac{\det\left[\mathbf{X}^T(i)\mathbf{X}(i)\right]}{\det(\mathbf{X}^T\mathbf{X})} = \frac{(1-h_i)}{\left\{\left[(n-p-1)/(n-p)\right] + \left[e_i^{*2}/(n-p)\right]\right\}^p}, \quad (2.48)$$

which is just (2.38) multiplied by $(1-h_i)^2$. We prefer (2.38) because no substitution for e_i^2 is required.

A similar result for the variances of the fit, $\hat{y}_i$, compares the ratio of $\Sigma_{k\neq i} e_k^2 h_{ik}^2$ and $s^2 h_i$ giving, after some manipulation,

$$\frac{s^2(i)(1-h_i)}{s^2} = \frac{1-h_i}{\left(\dfrac{n-p-1}{n-p} + \dfrac{e_i^{*2}}{n-p}\right)}, \tag{2.49}$$

which we note to be FVARATIO multiplied by $(1-h_i)^2$. This ratio can be related to some of the geometric procedures discussed below.

A Geometric View. In the previous sections we have examined several techniques for diagnosing those observations that are influential in the determination of various regression outputs. We have seen that key roles are played in these diagnostic techniques by the elements of the hat matrix **H**, especially its diagonal elements, the h_i, and by the residuals, the e_i. The former elements convey information from the **X** matrix, while the latter also introduce information from the response vector, **y**. A geometric way of viewing this interrelationship is offered by adjoining the **y** vector to the **X** matrix to form a matrix $\mathbf{Z} \equiv [\mathbf{X}\mathbf{y}]$, consisting of $p+1$ columns. We can think of each row of **Z** as an observation in a $p+1$ dimensional space and search for "outlying" observations.

In this situation, it is natural to think of Wilks' Λ statistic [Rao (1973)] for testing the differences in mean between two populations. Here one such population is represented by the ith observation and the second by the rest of the data. If we let $\tilde{\mathbf{Z}}$ denote the centered (by $\bar{\mathbf{z}}$) **Z** matrix, then the statistic is

$$\Lambda(\tilde{\mathbf{z}}_i) = \frac{\det\left(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}} - (n-1)\bar{\tilde{\mathbf{z}}}^T(i)\bar{\tilde{\mathbf{z}}}(i) - \tilde{\mathbf{z}}_i^T\tilde{\mathbf{z}}_i\right)}{\det(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})},$$

where $\bar{\tilde{\mathbf{z}}}(i)$ is the p-vector (row) of column means of $\tilde{\mathbf{Z}}(i)$.

As part of our discussion of the hat matrix in Appendix 2A, we show that

$$\Lambda(\tilde{\mathbf{x}}_i) = \frac{n}{n-1}(1-h_i), \tag{2.50}$$

and a simple application of the formulas for adding a column to a matrix [Rao (1973), p. 33] shows that

$$\Lambda(\tilde{\mathbf{z}}_i) = \left(\frac{n}{n-1}\right)(1-h_i)\left[1 + \frac{e_i^{*2}}{(n-p-1)}\right]^{-1}. \tag{2.51}$$

This index is again seen to be composed of the basic building blocks, h_i, and the studentized residuals, e_i^*, and is similar (in the case of a single observation in one group) to (2.49). Small values of (2.51) would indicate possible discrepant observations.

If we are willing to assume, for purposes of guidance, that $\tilde{\mathbf{Z}}$ consists of n independent samples from a p-dimensional Gaussian distribution, then $\Lambda(\tilde{\mathbf{z}}_i)$ can be easily related to the F-statistic by

$$\left(\frac{n-p-1}{p}\right)\frac{1-\Lambda(\tilde{\mathbf{z}}_i)}{\Lambda(\tilde{\mathbf{z}}_i)} \sim F_{p,n-p-1}. \tag{2.52}$$

In place of $\Lambda(\tilde{\mathbf{z}}_i)$ we could have used the Mahalanobis distance between one row and the mean of the rest; that is,

$$M(\tilde{\mathbf{z}}_i) = (n-2)\big(\tilde{\mathbf{z}}_i - \bar{\tilde{\mathbf{z}}}(i)\big)\big(\tilde{\tilde{\mathbf{Z}}}^T(i)\tilde{\tilde{\mathbf{Z}}}(i)\big)^{-1}\big(\tilde{\mathbf{z}}_i - \bar{\tilde{\mathbf{z}}}(i)\big)^T, \tag{2.53}$$

where $\tilde{\tilde{\mathbf{Z}}}(i)$ is $\tilde{\mathbf{Z}}(i)$ centered by $\bar{\tilde{\mathbf{z}}}(i)$. This is seen by noting that Λ and M are simply related by

$$\frac{1-\Lambda}{\Lambda} = \frac{(n-1)M}{(n-2)n}. \tag{2.54}$$

However, $\Lambda(\tilde{\mathbf{x}}_i)$ has a more direct relationship to h_i and its computation is somewhat easier when, later on, we consider removing more than one observation at a time.

The major disadvantage of diagnostic approaches based on $\mathbf{Z}$ is that the special nature of $\mathbf{y}$ in the regression context is ignored (except when $\mathbf{X}$ is considered as fixed in the distribution of diagnostics based on $\mathbf{Z}$). The close parallel of this approach to that of the covariance comparisons as given in (2.48) and (2.49) suggests, however, that computations based on $\mathbf{Z}$ will prove useful as well.

Criteria for Influential Observations. In interpreting the results of each of the previously described diagnostic techniques, a problem naturally arises in determining when a particular measure of leverage or influence is large enough to be worthy of further notice. When, for example, is a hat-matrix diagonal large enough to indicate a point of leverage, or a DFBETA an influential point? As with all empirical procedures, this question is ultimately answered by judgment and intuition in choosing reasonable cutoffs most suitable for the problem at hand, guided wherever possible by statistical theory. There are at least three sources of information for determining such cutoffs that seem useful: external

scaling, internal scaling, and gaps. Elasticities, such as $(\partial b_j(w_i)/\partial w_i)(w_i/b_j)$, and approximations to them like $(b_j - b_j(i))/b_j$, may also be useful in specific applications, but will not be pursued here.

External Scaling. External scaling denotes cutoff values determined by recourse to statistical theory. Each of the t-like diagnostics RSTUDENT, DFBETAS, and DFFITS, for example, has been scaled by an appropriate estimated standard error, which, under the Gaussian assumption, is stochastically independent of the given diagnostic. As such, it is natural to say, at least to a first approximation, that any of the diagnostic measures is large if its value exceeds two in magnitude. Such a procedure defines what we call an *absolute cutoff*, and it is most useful in determining cutoff values for RSTUDENT, since this diagnostic is less directly dependent on the sample size. Absolute cutoffs, however, are also relevant to determining extreme values for the diagnostics DFBETAS and DFFITS, even though these measures do depend directly on the sample size, since it would be most unusual for the removal of a single observation from a sample of 100 or more to result in a change in any estimated statistic by two or more standard errors. By way of contrast, there can be no absolute cutoffs for the hat-matrix diagonals h_i or for COVRATIO, since there is no natural standard-error scaling for these diagnostics.

While the preceding absolute cutoffs are of use in providing a stringent criterion that does not depend directly on the sample size n, there are many cases in which it is useful to have a cutoff that would tend to expose approximately the same proportion of potentially influential observations, regardless of sample size. Such a measure defines what we call a *size-adjusted cutoff*. In view of (2.7) and (2.9) a size-adjusted cutoff for DFBETAS is readily calculated as $2/\sqrt{n}$. Similarly, a size-adjusted cutoff for DFFITS is possible, for we recall from (2.19) that a perfectly balanced design matrix $\mathbf{X}$ would have $h_i = p/n$ for all i, and hence [see (2.11)],

$$\text{DFFITS}_i = \left(\frac{p}{n-p}\right)^{1/2} e_i^*.$$

A convenient size-adjusted cutoff in this case would be $2\sqrt{p/n}$, which accounts both for the sample size n and the fact that DFFITS_i increases as p does. In effect, then, the perfectly balanced case acts as a standard from which this measure indicates sizable departures. As we have noted above, the only cutoffs relevant to the hat-matrix diagonals h_i and COVRATIO are the size-adjusted cutoffs $2p/n$ and $1 \pm 3(p/n)$, respectively.

Both absolute and size-adjusted cutoffs have practical value, but the relation between them becomes particularly important for large data sets.

In this case, it is unlikely that the deletion of any single observation can result in large values for |DFBETAS| or |DFFITS|; that is, when n is large there are not likely to be any observations that are influential in the absolute sense. However, it is still extremely useful to discover those observations that are most strongly influential in relation to the others, and the size-adjusted cutoffs provide a convenient means for doing this.

Internal Scaling. Internal scaling defines extreme values of a diagnostic measure relative to the "weight of the evidence" provided by the given diagnostic series itself. The calculation of each deletion diagnostic results in a series of n values. The hat-matrix diagonals, for example, form a set of size n, as do DFFIT and the p series of DFBETA. Following Tukey (1977) we compute the interquartile range $\tilde{s}$ for each series and indicate as extreme those values that exceed $(7/2)\tilde{s}$. If these diagnostics were Gaussian this would occur less than 0.1% of the time. Thus, these limits can be viewed as a convenient point of departure in the absence of a more exact distribution theory. The use of an interquartile range in this context provides a more robust estimate of spread than would the standard deviation when the series are non-Gaussian, particularly in instances where the underlying distribution is heavy tailed.[12]

Gaps. With either internal or external scaling, we are always alerted when a noticeable gap appears in the series of a diagnostic measure; that is, when one or more values of the diagnostic measure show themselves to be singularly different from the rest. The question of deciding when a gap is worthy of notice is even more difficult than deriving the previous cutoffs. Our experience with the many data sets examined in the course of our research, however, shows that in nearly every instance a large majority of the elements of a diagnostic series bunches in the middle, while the tails frequently contain small fractions of observations clearly detached from the remainder.

It is important to note that, in any of these approaches to scaling, we face the problems associated with extreme values, multiple tests, and multiple comparisons. Bonferroni-type bounds can be useful for small data sets or for situations where only few diagnostics need to be examined because the rest have been excluded on other grounds. Until more is known about the issue, we suggest a cautious approach to the use of the

[12] For further discussion of appropriate measures of spread for non-Gaussian data, see Mosteller and Tukey (1977).

diagnostics, but not so cautious that we remain ignorant of the potentially damaging effects of highly influential data.

Partial-Regression Leverage Plots. Simple two-variable regression scatter-plots (like the stylized examples in Exhibit 2.1*e* and *f*) contain much diagnostic information about residuals and leverage and, in addition, provide guidance about influential subsets of data that might escape detection through the use of single-row techniques.

It is natural to ask if a similar tool exists for multiple regression, and this leads to the *partial-regression leverage plot*. This graphical device can be motivated as follows. Let $\mathbf{X}[k]$ be the $n \times (p-1)$ matrix formed from the data matrix, $\mathbf{X}$, by removing its kth column, $\mathbf{X}_k$. Further let $\mathbf{u}_k$ and $\mathbf{v}_k$, respectively, be the residuals that result from regressing $\mathbf{y}$ and $\mathbf{X}_k$ on $\mathbf{X}[k]$. As is well known, the kth regression coefficient of a multiple regression of $\mathbf{y}$ on $\mathbf{X}$ can be determined from the simple two-variate regression of $\mathbf{u}_k$ on $\mathbf{v}_k$. We define, then, the partial-regression leverage plot for b_k as a scatter plot of the $\mathbf{u}_k$ against the $\mathbf{v}_k$ along with their simple linear-regression line. The residuals from this regression line are, of course, just the residuals from the multiple regression of $\mathbf{y}$ on $\mathbf{X}$, and the slope is b_k, the multiple-regression estimate of β_k. Also, the simple correlation between $\mathbf{u}_k$ and $\mathbf{v}_k$ is equal to the partial correlation between $\mathbf{y}$ and $\mathbf{X}_k$ in the multiple regression.

We feel that these plots are an important part of regression diagnostics and that they should supplant the traditional plots of residuals against explanatory variables. Needless to say, however, partial-regression leverage plots cannot tell us everything. Certain types of multivariate influential data can be overlooked and the influence of the leverage points detected in the plot will sometimes be difficult to quantify. The computational details for these plots are discussed by Mosteller and Tukey (1977) who show that the $\mathbf{u}_k$ are equal to $b_k\mathbf{v}_k + \mathbf{e}$, where $\mathbf{e}$ is the vector of residuals from the multiple regression. This fact saves considerable computational effort.

The $\mathbf{v}_k$ have another interesting interpretation. Let $h_{ij}[k]$ denote the elements of the hat matrix for the regression of $\mathbf{y}$ on all of the explanatory variables except $\mathbf{X}_k$. Then the elements of the hat matrix for the full regression are

$$h_{ij} = h_{ij}[k] + \frac{v_{k,i}v_{k,j}}{\sum_{l=1}^{n} v_{k,l}^2}, \tag{2.55}$$

where $v_{k,i}$ denotes the ith component of the vector $\mathbf{v}_k$. This expression can be usefully compared with (2.21) for regression through the origin. Thus the $\mathbf{v}_k$ are closely related to the partial leverage added to $h_{ij}[k]$ by the addition of $\mathbf{X}_k$ to the regression.

Multiple-Row Effects

In the preceding discussion, we have presented various diagnostic techniques for identifying influential observations that have been based on the deletion or alteration of a single row. While such techniques can satisfactorily identify influential observations much of the time, they will not always be successful. We have already seen, for example, in the simple case presented in Exhibit 2.1*f* that one outlier can mask the effect of another. It is necessary, therefore, to develop techniques that examine the potentially influential effects of subsets or groups of observations. We turn shortly to several multiple-row techniques that tend to avoid the effects of masking and that have a better chance of isolating influential subsets in the data.

Before doing this, however, we must mention an inherent problem in delimiting influential subsets of the data, namely, when to stop—with subsets of size two, three, or more? Clearly, unusual observations can only be recognized relative to the bulk of the remaining data that are considered to be typical, and we must select an initial base subset of observations to serve this purpose. But how is this subset to be found? One straightforward approach would be to consider those observations that do not appear exceptional by any of the single-row measures discussed above. Of course, we could always be fooled, as in the example of Exhibit 2.1*f*, into including some discrepant observations in this base subset, but this would be minimized if we used low cutoffs, such as relaxing our size-adjusted cutoff levels to 90% or less instead of holding to the more conventional 95% level. We could also remove exceptional observations noticed in the partial-regression leverage plots. Some of the following procedures are less dependent on a base subset than others, but it cannot be avoided entirely, for the boundary between the typical and the unusual is inherently vague. We denote by B^* (of size m^*) the largest subset of potentially influential observations that we wish to consider. The complement of B^* is the base subset of observations defined to be typical.

We follow the same general outline as before and discuss deletion, residuals, differentiation, and geometric approaches in the multiple-row context.

Deletion. A natural multiple-row generalization of (2.4) would be to examine the larger values of

$$\frac{|b_j - b_j(D_m)|}{\text{scale}}, \tag{2.56}$$

for $j = 1,\ldots,p$ and $m = 2, 3, 4$, and so on, and where "scale" indicates some

appropriate measure of standard error. Here D_m is a set (of size m) of indexes of the rows to be deleted. If fitted values are of interest, then the appropriate measure becomes

$$\frac{|\mathbf{x}_k[\mathbf{b}-\mathbf{b}(D_m)]|}{\text{scale}}, \tag{2.57}$$

for $k=1,\ldots,n$. Although computational formulas exist for these quantities [Bingham (1977)], the cost is great and we feel most of the benefits can be obtained more simply.

To avoid the consideration of p quantities in (2.56) or n quantities in (2.57), squared norms, such as

$$[\mathbf{b}-\mathbf{b}(D_m)]^T[\mathbf{b}-\mathbf{b}(D_m)] \tag{2.58}$$

or

$$[\mathbf{b}-\mathbf{b}(D_m)]^T\mathbf{X}^T\mathbf{X}[\mathbf{b}-\mathbf{b}(D_m)] \tag{2.59}$$

can be considered as summary measures. Since we are often most interested in changes in fit that occur for the data points remaining after deletion, (2.59) can be modified to

$$\text{MDFFIT} \equiv [\mathbf{b}-\mathbf{b}(D_m)]^T\mathbf{X}^T(D_m)\mathbf{X}(D_m)[\mathbf{b}-\mathbf{b}(D_m)]. \tag{2.60}$$

Bingham (1977) shows (2.60) can also be expressed as

$$\mathbf{e}_{D_m}^T\mathbf{X}_{D_m}[\mathbf{X}^T(D_m)\mathbf{X}(D_m)]^{-1}\mathbf{X}_{D_m}^T\mathbf{e}_{D_m}, \tag{2.61}$$

where $\mathbf{e}$ is the column vector of least-squares residuals and where D_m, used as a subscript, denotes a matrix or vector with rows whose indexes are contained in D_m. Because of (2.61) MDFFIT can be computed at lower cost than (2.59). Unfortunately, even (2.61) is computationally expensive when m exceeds about 20 observations. Some inequalities, however, are available for MDFFIT which may ease these computational problems. More details are provided at the end of Appendix 2B.

For larger data sets, a stepwise approach is available that can provide useful information at low cost. This method begins for $m=2$ by using the two largest |DFFIT| (or |DFFITS|) to form $D_2^{(1)}$. If the two largest values

of

$$|\mathbf{x}_k[\mathbf{b}-\mathbf{b}(D_2^{(1)})]| \tag{2.62}$$

do not have their indexes k contained in $D_2^{(1)}$, a set $D_2^{(2)}$ is formed consisting of the indexes for the two largest. This procedure is iterated until a set D_2 is found with indexes coinciding with the two largest values of (2.62). The resulting statistic is designated SMDFFIT.

For $m=3$, a starting set $D_3^{(1)}$ is found by using the three largest values of (2.62) from the final iteration for $m=2$. Once the starting set is found the iteration proceeds as for $m=2$. The overall process continues for $m=4$, 5, and so on. An alternative approach is to use the m largest values of |DFFIT| to start the iterations for each value of m. Different starting sets can lead to different final results.

This stepwise approach is motivated by the idea that the fitted values most sensitive to deletion should be those which correspond to the deleted observations because no attempt is being made to fit these points. Since (2.14) does not hold in general when two or more points are deleted, the stepwise process attempts to find a specific set for each m where it does hold.

We conclude our study of multiple-row deletion by generalizing the covariance ratio to a deletion set D_m; namely,

$$\text{COVRATIO}(D_m) \equiv \frac{\det s^2(D_m)\left[\mathbf{X}^T(D_m)\mathbf{X}(D_m)\right]^{-1}}{\det s^2(\mathbf{X}^T\mathbf{X})^{-1}}. \tag{2.63}$$

Computation of this ratio is facilitated by the fact that

$$\frac{\det\left[\mathbf{X}^T(D_m)\mathbf{X}(D_m)\right]}{\det(\mathbf{X}^T\mathbf{X})} = \det(\mathbf{I}-\mathbf{H})_{D_m}, \tag{2.64}$$

where $(\mathbf{I}-\mathbf{H})_{D_m}$ stands for the submatrix formed by considering only the rows and columns of $\mathbf{I}-\mathbf{H}$ that are contained in D_m. FVARATIO also can be generalized.[13]

Studentized Residuals and Dummy Variables. The single-row studentized residual given in (2.26) is readily extended to deletions of more than one row at a time. Instead of adding just one dummy variate with a unity in row i and zeros elsewhere, we add many such dummies, each with its unity

[13] It is easily seen that equations (2.48) and (2.49) can also be generalized.

only in the row to be deleted. In the extreme, we could add n such variables, one for each row. This leads to a singular problem which can, in fact, be studied. However, we assume that no more than $n-p$ columns of dummy variates are to be added.

Once the subset of dummy columns to be added has been decided on, a problem that we turn to below, it is natural to make use of standard regression selection techniques to decide which, if any, of these dummy variables should be retained. Each dummy variable that is retained indicates that its corresponding row warrants special attention, just as we saw that the studentized residual calls attention to a single observation. The advantage here is that several rows can be considered simultaneously and we have a chance to overcome the masking situation in Exhibit 2.1*f*.

There are no clear-cut means for selecting the set of dummy variables to be added. As already noted, we could use the previously described single-row techniques along with partial-regression leverage plots to determine a starting subset of potentially influential observations. Rather generally, however, the computational efficiency of some of these selection algorithms allows this starting subset to be chosen quite large.

To test any particular subset D_m of dummy variables a generalization of (2.27) is available. For example, we could consider

$$\text{RESRATIO} \equiv \frac{[\text{SSR}(\text{no dummies}) - \text{SSR}(D_m \text{ dummies used})]/m}{\text{SSR}(D_m \text{ dummies used})/(n-p-m)}, \tag{2.65}$$

which is distributed as $F_{m,n-p-m}$ if the appropriate probability assumptions hold. For further details see Gentleman and Wilk (1975).

The use of stepwise regression has been considered as a solution to this problem by Mickey, Dunn, and Clark (1967). The well-known difficulties of stepwise regression arise in this context, and, in general, it is best to avoid attempting to discover the model (i.e., explanatory variables) and influential points at the same time. Thus one must first choose a set of explanatory variables and stay with them while the dummy variables are selected. Of course, this process may be iterated and, if some observations are deleted, a new stepwise regression on the explanatory variable set should be performed. Stepwise regression also clearly fails to consider all possible combinations of the dummy variables and can therefore miss influential points when more than one is present.

A natural alternative to stepwise regression is to consider all-possible-subsets regression.[14] The computational costs are higher and

[14] See Furnival and Wilson (1974).

more care must be taken in choosing the starting subset of dummy variables. Wood (1973) has suggested using partial-residual plots[15] to find an initial subset which is subjected in turn to the C_p selection technique developed in Mallows (1973b) in order to find which dummy variables are to be retained. We think this method is appealing, especially if partial-regression leverage plots are combined with the methods discussed earlier in this chapter as an aid to finding the initial subset of dummies. Computational costs will tend to limit the practical size of this subset to 20 or fewer dummy variates.

The use of dummy variates has considerable appeal but the single-row analogue, the studentized residual, is, as we have seen, clearly not adequate for finding influential data points. This criticism extends to the dummy-variable approach because the use of sums of squares of residuals fails to give adequate weight to the structure and leverage of the explanatory-variable data.

The deletion methods discussed above provide one way to deal with this failure. Another is to realize that $\mathbf{I}-\mathbf{H}$ is proportional to the covariance matrix of the least-squares residuals. A straightforward argument using (2.64) shows that

$$1-h_i(D_m)=\frac{\det(\mathbf{I}-\mathbf{H})_{D_m,i}}{\det(\mathbf{I}-\mathbf{H})_{D_m}}, \tag{2.66}$$

where the numerator submatrix of $(\mathbf{I}-\mathbf{H})$ contains the ith row and column of $\mathbf{I}-\mathbf{H}$ in addition to the rows and columns in D_m. When this is specialized to a single deleted row, $k\neq i$, we obtain

$$\begin{aligned}1-h_i(k)&=\frac{(1-h_i)(1-h_k)-h_{ik}^2}{1-h_k}\\&=(1-h_i)\left[1-\operatorname{cor}^2(e_i,e_k)\right].\end{aligned} \tag{2.67}$$

This means that $h_i(k)$ can be large when the magnitude of the correlation between e_i and e_k is large. Thus useful clues about subsets of leverage points can be provided by looking at large diagonal elements of $\mathbf{H}$ and at the large residual correlations. This is an example of the direct use of the off-diagonal elements of $\mathbf{H}$, elements implicitly involved in most multiple-row procedures. This is further exemplified in the next two sections.

Differentiation. Generalizing the single-row differentiation techniques to multiple-row deletion is straightforward. Instead of altering the weight, w_i, attached to only one observation, we now consider a diagonal weight

[15] On these plots, see Larson and McCleary (1972).

matrix $\mathbf{W}$, with diagonal elements $(w_1, w_2, \ldots, w_n) \equiv \mathbf{w}$, and define $\mathbf{b}(\mathbf{w}) \equiv (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$. This $\mathbf{b}(\mathbf{w})$ is a vector-valued function of $\mathbf{w}$ whose first partial derivatives evaluated at $\mathbf{w} = \boldsymbol{\iota}$ (the vector of ones) are

$$\nabla\mathbf{b}(\boldsymbol{\iota}) = \mathbf{C}\mathbf{E}, \tag{2.68}$$

where ∇ is the standard gradient operator, $\mathbf{C}$ is defined in (2.3), and $\mathbf{E} = \operatorname{diag}(e_1, \ldots, e_n)$. If we are interested in fitted values, this becomes

$$\mathbf{X}\nabla\mathbf{b}(\boldsymbol{\iota}) = \mathbf{H}\mathbf{E}. \tag{2.69}$$

Our concern is with subsets of observations that have a large influence. One way to identify such subsets is to consider the directional derivatives $\nabla\mathbf{b}(\boldsymbol{\iota})\boldsymbol{l}$ where $\boldsymbol{l}$ is a column vector of unit length with nonzero entries in rows with indexes in D_m, that is, the rows to be perturbed. For a fixed m, the indexes corresponding to the nonzero entries in those $\boldsymbol{l}$ which give large values of

$$\boldsymbol{l}^T\nabla\mathbf{b}^T(\boldsymbol{\iota})\mathbf{A}\nabla\mathbf{b}(\boldsymbol{\iota})\boldsymbol{l} \tag{2.70}$$

would be of interest. The matrix $\mathbf{A}$ is generally $\mathbf{I}$, $\mathbf{X}^T\mathbf{X}$, or $\mathbf{X}^T(D_m)\mathbf{X}(D_m)$.

These $\boldsymbol{l}$ vectors are just the eigenvectors corresponding to largest eigenvalues of the matrix

$$\left[\nabla\mathbf{b}^T(\boldsymbol{\iota})\mathbf{A}\nabla\mathbf{b}(\boldsymbol{\iota})\right]_{D_m}. \tag{2.71}$$

When $\mathbf{A} = \mathbf{X}^T\mathbf{X}$, (2.71) is just the matrix whose elements are $h_{ij}e_ie_j$, with $i, j \in D_m$. While the foregoing procedure is conceptually straightforward, it has the practical drawback that, computationally, finding these eigenvectors is expensive. We therefore explore two less costly simplifications.

In the first simplification we place equal weight on all the rows of interest, and consider the effect of an infinitesimal perturbation of that single weight. This is equivalent to using a particular directional derivative, $\boldsymbol{l}^*$, that has all of its nonzero entries equal. When $\mathbf{A}$ is $\mathbf{X}^T\mathbf{X}$, this gives

$$\boldsymbol{l}^{*T}\nabla\mathbf{b}^T(\boldsymbol{\iota})\mathbf{X}^T\mathbf{X}\nabla\mathbf{b}(\boldsymbol{\iota})\boldsymbol{l}^* = \sum_{i,j \in D_m} h_{ij}e_ie_j. \tag{2.72}$$

Much less computational effort is required to find the large values of (2.72) than to compute the eigenvectors for (2.71). A more complete discussion of this issue is contained in Appendix 2B. The expression $\mathbf{b} - \mathbf{b}(i)$ is an approximation to the ith column of $\nabla\mathbf{b}(\boldsymbol{\iota})$, and could be used instead in the

preceding discussion. In this case (2.72) becomes

$$\sum_{i,j\in D_m} h_{ij}\frac{e_i e_i}{(1-h_i)(1-h_j)} \equiv \text{MEWDFFIT}. \tag{2.73}$$

In the second simplification, we use a stepwise approach for large data sets, employed in the same manner as (2.62), using the statistic

$$\left|\mathbf{x}_k \sum_{i\in D_m} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T e_i\right| = \left|\sum_{i\in D_m} h_{ki}e_i\right|. \tag{2.74}$$

Geometric Approaches. Wilks' Λ statistic generalizes to the multiple-row situation quite readily and is useful for discovering groups of outliers. This is particularly interesting when the observations cannot be grouped on the basis of prior knowledge (e.g., time) or when there is prior knowledge but unexpected groupings occur.

The generalization goes as follows. Let $\boldsymbol{l}_1$ be an $n\times 1$ vector consisting of ones for rows contained in D_m and zeros elsewhere and $\boldsymbol{l}_2=\iota-\boldsymbol{l}_1$. The relevant Λ statistic for this case is [Rao (1973), p. 570]

$$\Lambda(D_m)=\frac{\det\left[\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}-(1/m)\tilde{\mathbf{Z}}^T\boldsymbol{l}_1\boldsymbol{l}_1{}^T\tilde{\mathbf{Z}}-(1/(n-m))\tilde{\mathbf{Z}}^T\boldsymbol{l}_2\boldsymbol{l}_2^T\tilde{\mathbf{Z}}\right]}{\det(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})}.$$

Using an argument similar to that in Appendix 2A, this statistic reduces to

$$\Lambda(D_m)=1-\frac{n}{m(n-m)}(\boldsymbol{l}_1^T\tilde{\mathbf{P}}\boldsymbol{l}_1), \tag{2.75}$$

where $\tilde{\mathbf{P}}\equiv\tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T$. Thus $\Lambda(D_m)$ is directly related to sums of elements of a matrix, $\tilde{\mathbf{P}}$, ($\tilde{\mathbf{H}}$ if $\tilde{\mathbf{Z}}$ is replaced by $\tilde{\mathbf{X}}$) and, as we show in Appendix 2B, this greatly simplifies computation.

To use Λ we examine the smaller values for each $m=1$, 2, and so on. If we assume for guidance that the rows of $\tilde{\mathbf{Z}}$ are independent samples from a p-variate Gaussian distribution, then

$$\left(\frac{n-p-1}{p}\right)\left[\frac{1-\Lambda(D_m)}{\Lambda(D_m)}\right]\sim F_{p,n-p-1}. \tag{2.76}$$

This is only approximate, since we are interested in extreme values. It would be even better to know the distribution of Λ conditional on $\mathbf{X}$, but

this remains an open problem. More than just the smallest value of Λ should be examined for each m, since there may be several significant groups. Gaps in the values of Λ are also usually worth noting.

Andrews and Pregibon (1978) have proposed another method based on $\mathbf{Z}$. They consider the statistic

$$Q(D_m) = \frac{\det\left[\mathbf{Z}^T(D_m)\mathbf{Z}(D_m)\right]}{\det(\mathbf{Z}^T\mathbf{Z})}$$

$$= \frac{(n-p-m)s^2(D_m)\det\left[\mathbf{X}^T(D_m)\mathbf{X}(D_m)\right]}{(n-p)s^2\det(\mathbf{X}^T\mathbf{X})}, \tag{2.77}$$

which relates to (2.49) and (2.51) for $m=1$. The idea is to ascertain the change in volume (measured by the determinant of $\mathbf{Z}^T\mathbf{Z}$) caused by the deletion of the rows in D_m. If $\tilde{\mathbf{Z}}$ instead of $\mathbf{Z}$ had been used, Q becomes another form of Wilks' Λ statistic where there are $m+1$ groups: one for each row in D_m and one group for all the remaining rows.

Computationally, Q is about the same order as MDFFIT and considerably more complicated than Λ. However, Andrews and Pregibon have succeeded in developing a distribution theory for Q when $\mathbf{y}$ is Gaussian and $\mathbf{X}$ is fixed. While useful only for n of modest size, it does provide some significance levels for finding sets of outliers.

Both Λ and Q are computationally feasible for $m \leqslant 20$. A stepwise approach based on the Mahalanobis distance and the ideas of robust covariance [Devlin, Gnanadesikan, and Kettenring (1975)] can be used for larger subsets. The philosophy is similar to that developed for (2.62) and (2.74). If we think the points in D are outliers, it is reasonable to remove them from our estimate of the covariance and means of the columns of $\tilde{\mathbf{Z}}$ by computing $\tilde{\tilde{\mathbf{Z}}}^T(D)\tilde{\tilde{\mathbf{Z}}}(D)$ and $\bar{\tilde{\mathbf{z}}}(D)$. The distance from any row $\tilde{\mathbf{z}}_i$ to $\bar{\tilde{\mathbf{z}}}(D)$ is then measured by

$$M(i,D) = (n-2)\left[\tilde{\mathbf{z}}_i - \bar{\tilde{\mathbf{z}}}(D)\right]\left[\tilde{\tilde{\mathbf{Z}}}^T(D)\tilde{\tilde{\mathbf{Z}}}(D)\right]^{-1}\left[\tilde{\mathbf{z}}_i - \bar{\tilde{\mathbf{z}}}(D)\right]^T. \tag{2.78}$$

The starting set $D_2^{(1)}$ consists of the rows corresponding to the two largest values of the single-row Mahalanobis distance $M(\tilde{\mathbf{z}}_i)$. $D_2^{(2)}$ consists of the indexes of the two largest values of $M(i, D_2^{(1)})$. If $D_2^{(2)} = D_2^{(1)}$, we stop. If not we iterate with $D_2^{(k+1)}$ consisting of the two largest values of $M(i, D_2^{(k)})$, and the process stops when $D_2^{(k+1)} = D_2^{(k)}$.

Final Comments

The multiple-row techniques presented here form a subset of the possible procedures that could be devised. Our choices have been made on the