Influential data points can affect the results of a regression analysis; for example, the usual summary statistics and tests of significance may be misleading. The importance of regression diagnostics in detecting influential points is discussed, and five statistics are recommended for the applied researcher. The suggested diagnostics were used on a small dataset to detect an influential data point, and the effects were analyzed. Colinearity-based diagnostics also are discussed and illustrated on the same dataset. The non-robustness of the least squares estimates in the presence of influential points is emphasized. Diagnostics for multiple influential points, multivariate regression, multicollinearity, nonlinear regression, and other multivariate procedures also are discussed. Index terms: Andrew-Pregibon measure, colinearity, Cook's distance, covariance ratio, influential observations, measurement error, partial residual plot, regression diagnostics.

Explanation of the relationships among variables is a major goal of the behavioral and social sciences. To achieve this goal, relevant variables and constructs must be discovered and measured on reasonably precise scales. Because many behavioral variables do not lend themselves to exact measurement, establishment of reliable and valid measures for these variables is a continuing challenge facing behavioral scientists.

Common approaches to the study of relationships among variables include correlation and regression models. For these models to be useful, the variables must be measured with as little error as possible because errors in measurement tend to attenuate or distort the relationships among variables, which causes serious difficulties in the interpretation of results. Consequently, avoiding additional distortions that arise from influential data points and outliers gains added importance in the use of these models in the behavioral sciences.

Observed variables include errors only when a model can be specified for the variables that distinguishes between underlying unobservable and observable components of measurement errors. Regression models have extensive psychometric problems. For example, Lord & Novick (1968) and Nunnally (1967) include in-depth discussions of corrections for attenuation, restriction of range, unreliability, measurement errors, and other errors. Although awareness of the psychometric problems is vital, it is also important to recognize the estimation bias arising from discrepant data points and the important role that diagnostics play in uncovering such points. Although the concern here is less with the psychometric issues and more with the study of regression diagnostics, an integrative approach is used because problems such as measurement errors are common to both concerns. Only a full understanding of the importance of both kinds of issues can lead to acceptable models for the social and behavioral sciences.

The importance of influential data points and outliers on estimates calculated in a linear regression model has generated such a large body of work that a new subfield—regression diagnostics—has developed. With some recent exceptions noted below, this work has received little attention in statistical analysis books used by social scientists or in social science journals. For ex-
ample, Pedhazur (1982) only briefly discusses outliers and influential points, even though popular statistical software provides several diagnostics [e.g., BMDP (Dixon, 1984), Minitab (Ryan, Joiner, & Ryan, 1991), SAS (SAS Institute, Inc., 1982), SPSS (SPSS Inc., 1988), and SYSTAT (Wilkinson, 1992)]. Recent reviews somewhat narrower in scope than this review include Cook & Weisberg (1982b), Stevens (1984), and Bollen & Jackman (1985). Most recent applied statistics books (e.g., Darlington, 1990) also contain at least some discussion of regression diagnostics. Specialized books by Chatterjee & Hadi (1988), Atkinson (1985), Cook & Weisberg (1982a), and Belsley, Kuh, & Welsch (1980) deal exclusively with regression diagnostics, influential observations, and outliers. For useful and substantive applications of regression diagnostics in the social sciences, see Bollen & Jackman (1985) and Chatterjee & Wiseman (1984). The present paper is intended to be broader in scope and less technical in its presentation. The concepts behind the statistics are emphasized at the cost of mathematical rigor. Formal tests of hypotheses are not emphasized, because regression diagnostics are intended as exploratory data analysis.

**Regression Models**

This paper considers the general linear regression model,

\[
y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i,\]

or \( y = x\beta + \varepsilon \) in matrix notation, where \( y = (y_1, \ldots, y_n)^T \) is a \( n \times 1 \) column vector of \( n \) observations of a stochastic response (dependent or criterion) variable \( y \), and \( T \) denotes the transpose. For the \( i \)th observation, \( x_{i1}, x_{i2}, \ldots, x_{ik} \) are the observed values of \( k \) regressor (independent or predictor) variables \( X_1, X_2, \ldots, X_k \). \( \beta = (\beta_0, \ldots, \beta_k)^T \) is a \( (k+1) \times 1 \) column vector of unknown constants, \( \varepsilon = (e_1, \ldots, e_n)^T \) is a \( n \times 1 \) column vector of random errors, and \( x \) is a \( n \times (k+1) \) matrix with its first column consisting of 1s, and the remaining columns containing the regressor values \( x_{ij}, i = 1, 2, \ldots, n \) and \( j = 1, 2, \ldots, k \). As usual, the ordinary least squares (OLS) estimates are denoted by \( b = (b_0, \ldots, b_k)^T \), given by \( b = (x^T x)^{-1} x^T y \), which is obtained by minimizing \( \varepsilon^T \varepsilon \) with respect to \( \beta \). It is assumed that the random errors \( \varepsilon \) are independent with mean 0 and variance \( \sigma^2 \) (homoscedasticity). If the \( x \)s can be regarded as fixed constants, then the OLS estimates are optimal in the sense of being unbiased and having minimum variance among all linear estimators (i.e., the best linear unbiased estimator). The estimates of the slopes are given by \( b_0, b_1, \ldots, b_k \); the estimate of the standard error of the regression equation \( \sigma^2 \) is given by

\[
\hat{\sigma}^2 = \frac{(y - \hat{b})^T (y - \hat{b})}{n - k - 1}
\]

and the measure of fit by

\[
R^2 = \frac{\hat{y}^T \hat{y} - \bar{y}^T \bar{y}}{\hat{y}^T \hat{y}}
\]

where \( \hat{y} \) is the fitted value. The adjusted \( R^2 \) value is given by

\[
R_{adj}^2 = 1 - \left( \frac{n - 1}{n - k} \right) (1 - R^2)
\]

For purposes of inference beyond point estimation, such as tests of hypotheses, errors are often assumed to be normally distributed.

If either assumption—homoscedasticity or independence of errors—is not tenable, it is still theoretically possible to obtain optimal estimates of the \( \beta \)s using generalized least squares estimation, which minimizes \( \varepsilon^T \Sigma^{-1} \varepsilon \), where \( \Sigma \) is the \( n \times n \) variance-covariance matrix of \( \varepsilon \). If the observed values of the regressors are regarded as fixed constants, then the values of the regressors should be fixed prior to sampling, and repeated observations of the response variable should be made only for these fixed values of the regressors. In most behavioral studies, it is not feasible or appropriate to conduct experiments by fixing the regressors as known constants. Fortunately, the linear model and its estimation remain essentially unchanged if the conditional distributions

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/
of the $y_i$, given $x = \{x_i\}$, are independent with constant variance $\sigma^2$, and regressors are independent random variables with distributions that do not depend on the parameters $\beta$ or the constant variance $\sigma^2$. Even correlated errors present no additional problems if it can be assumed that the distribution of $x$ does not depend on $\beta$ or $\Sigma$.

Because the true values of these parameters are unknown, sample estimates provide the only adequate means of testing the validity of these assumptions in practice. This is usually achieved by examining plots of residuals (the observed value minus the fitted value) resulting from an estimated model, such as the plots of residuals against the predicted response or against each regressor variable. If the assumptions are not met, however, parameter estimates obtained from OLS either will be biased or have other non-optimal properties (such as lack of consistency). Least squares estimates will be nonoptimal, for example, if the regressor variables are stochastic but are correlated with the errors.

If the distributional assumptions stated above are not met for the stochastic variables, a suggested alternative to OLS estimation involves searching for additional variables that can be used to model the regressor variables and that satisfy the usual assumptions. These lead to two-stage least squares estimation and linear structural equations models that are outside the scope of the present review (see, e.g., Bollen, 1989; Judge, Griffith, Hill, & Lee, 1980; Malinvaud, 1970).

Models With Measurement Error

Errors in measuring the response variable $y$ do not require any special attention, because these errors are absorbed in the random errors of the regression model. Thus, the net effect of errors in $y$ is a larger standard error of the estimated regression coefficients. On the other hand, measurement errors in regressor variables produce error terms that are correlated with the regressors. Measurement error may affect statistical analysis because it can cause the probability distribution of observed data to differ from the distribution of the error-free data (see Cochran, 1972, Stefanski, 1985, and Chesher, 1991 for information on measurement error models). As noted above, the presence of measurement error leads to estimates of the regression coefficients that are nonoptimal (i.e., biased or inconsistent). The effect of this nonoptimality can be investigated using (1) the asymptotic approach, (2) the perturbation approach, and (3) the simulation approach.

In the asymptotic approach, which is popular in econometrics, large sample biases are calculated analytically and limiting values are sought (Chatterjee & Hadi, 1986). In the perturbation approach, the effects on the regression coefficients of perturbing the regressors by small amounts are studied analytically. This method provides an upper bound on the relative errors in the estimated coefficients (see Stewart, 1987). In the simulation approach, the structure of the errors in the regressor variables is simulated, and the effects are studied on the estimated regression coefficients (Chatterjee & Hadi, 1986). These three approaches require the researcher to make assumptions about the form of the distributions, the moments of the distributions, and/or the bounds on the errors in the variables. Refer to Chatterjee & Hadi (1988) for details of all three methods. Fuller (1987) and Heller (1987) provided specialized discussions for studying various aspects of measurement errors on the parameter estimates, and Bollen (1989) discussed modeling in the presence of these errors.

In theory, "instrumental variables" may provide a possible solution to the problem of measurement errors in the regression variables. Instrumental variables correlate very highly with the regressors but not with the measurement errors in the regressors or the random errors associated with the dependent variable. If such variables can be found, they can be used as regressor variables; alternatively, regressor variables could be modeled in terms of the instrumental variables. This last approach leads to the estimation of simultaneous equations by the two-stage least squares approach mentioned above (Judge et al., 1980). Clearly, the main
problems with this approach are identifying the instrumental variables and collecting data for them. When measurement errors are of concern, the researcher must decide on a strategy for dealing with them before using regression diagnostics.

**Cross-Validation**

Regression diagnostics are useful for detecting influential observations and may help the user select the proper statistical model. Cross-validation is another technique used in the social sciences to test the appropriateness of a model for a given set of observations. In cross-validation, a portion of the data is not used in developing the regression model (Picard & Berk, 1990; Picard & Cook, 1984). Three different ways to conduct cross-validation are suggested in the literature: (1) the split-sample method, (2) the hold-out method, and (3) the leave-one-out or jackknife method.

The first two approaches are very similar techniques in that they are both simple and intuitively appealing. In the split-sample method, 50% of the observations are used for estimation, and the remaining 50% are used for the validation of the model under consideration. This procedure may result in inefficient and/or inaccurate estimation and validation if the total number of observations is not large. In the hold-out method, a larger portion (e.g., 80%) of the data are used for estimation, and a smaller portion (e.g., 20%) are used for validation. The rationale for holding out a larger portion of the observations is that the accuracy of the cross-validation cannot be judged without precise estimates. However, even with this method, a small sample might result in imprecise estimation.

The jackknife method is an attempt to combine the strengths of the above two methods. The regression model is estimated with \((n - 1)\) observations, and the observation which is left out is predicted from the regression model obtained from the \((n - 1)\) observations. This process is repeated by bringing in the previously left out observation and leaving out a different observation. Thus, \(n\) predictions are obtained that can be used to judge the efficacy of the model. In this case, a balance is reached for both estimation and prediction. This idea also appears in some regression diagnostics, such as the externally studentized residual.

**Influential Observations and Regression Diagnostics**

A distinction must be made between outliers and influential points. Outliers are unusually extreme values in the response variable. Influential variables are extreme values in regressor variables that have a disproportionate effect on parameter estimates such as the slope, estimated standard error of the regression equation, \(R^2\), and so forth. Because each point in a regression space is defined by a combination of a response variable and a regressor variable, outliers may or may not be influential observations and influential variables may or may not be outliers. Figure 1 illustrates this in the context of a bivariate linear regression. In Figure 1a, A' is an outlier but not an influential data point; if A' is shifted to position A'', it becomes an influential point and not an outlier. In Figure 1b, A' is an influential point but not an outlier. If A' is moved to position A'', then it continues to be an influential point but also becomes an outlier. In the context of multiple regression, simple two-dimensional scatter plots cannot be used and the value of diagnostics becomes apparent.

If influential points are present, serious errors may be made in interpreting the regression model. In extreme cases, it is possible to conclude that a hypothesized relationship exists when in fact it does not, or to conclude that no relationship exists when it does. In Figure 1c, point A is highly influential—the regression line has a positive slope when A is included, but it has a horizontal line indicating no relationship when A is excluded. In Figure 1d, the inclusion of the influential point A yields a horizontal regression line, whereas its exclusion gives a line having positive slope.

Point A in Figure 1c is no longer an influential point if a curvilinear model is used. Thus,
a point is influential only with respect to a model. Before searching for influential points or outliers a model must be selected; the model must be intrinsically nonlinear and must be one which cannot be transformed easily to a linear model.

Model selections are driven by a priori conditions, such as theory or the experience of other researchers. Using outlier and influence analysis, it may be discovered that the model should be modified. However, in practice, there is never a “true” or “correct” model; therefore, influential points and model selections may depend on each other. In practice, a proper balance between model selection motivated by theory and model selection guided by data must be maintained. This paradigm of an iteration between theory and data analysis has been supported by many, including Box (1983).

The importance of various diagnostics does not diminish as the sample size increases. Although the illustrative datasets here are based on small samples, the relevance of regression diagnostics remains valid for samples of any size (e.g., Rousseeuw & LeRoy, 1987).

**Regressor-Based Measures**

To study the impact of \( x \) and \( y \) on \( b \), define a \( n \times n \) prediction matrix, given by \( H = x(x'x)^{-1}x' \). Denote the fitted values \( \hat{y} \), where \( \hat{y} = Hy \), and \( H \) maps the observed values into the predicted values. The diagonal elements of \( H \) are denoted by \( h_i \), and the off-diagonal elements are denoted by \( h_{ij} \). From \( h_i = x_i(x'x)^{-1}x_i' \), it follows that the fitted value \( \hat{y}_i \) depends on the \( i \)th observation through the value of \( h_i \). Because \( h_i \) is only a function of the observed values of explanatory variables or the design matrix, it is called a regressor-based diagnostic measure. The \( h \) values are such that \( 0 \leq h_i \leq 1 \), and they also are called leverages because they indicate how extreme the observed regressor values are. High values of \( h_i \) are indicative of influence; an \( h_i \) value greater than \( 2(k + 1)/n \) for a moderate sample size \( (k > 10, n - k > 50) \) usually indicates the presence of an influential point. For smaller \( k \) and \( n \), an \( h_i \) value greater than \( 3(k + 1)/n \) is sug-
gested by Velleman & Welsch (1981) as indicating an influential point. $h_i$ values also enter into the calculation of other diagnostics. The diagonal elements also give some information about the influence of other observations on $\hat{y}_i$, because $h_i = \sum_{j=1}^{n} h_{ij}$, so that no entry $h_{ij}$ can be larger in absolute value than $(h_i)^{1/2}$. In other words, the contribution of the $j$th observation on the $i$th parameter is bounded by $(h_i)^{1/2}$.

**Analysis of Residuals**

The OLS residuals, $e_i = y_i - \hat{y}_i$, are the primary means for detecting outliers. Although the magnitudes of the OLS residuals are difficult to judge directly, their scaled versions, called studentized residuals, allow for easier detection of outliers.

For the $i$th observation, an internally studentized residual is defined as

$$t_i = \frac{e_i}{s(1 - h_i)^{1/2}},$$

where $s$ is the estimate of the standard error of the regression equation given by

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - k - 1},$$

and the externally studentized residual is given by

$$t_i^* = \frac{e_i}{s_0(1 - h_i)^{1/2}},$$

where $s_0$ is the standard error of the regression equation obtained from deleting the $i$th observation. The internally studentized residual uses the standard error $s$, in the scaling, whereas the externally studentized residual uses the standard error $s_0$ without the $i$th data point. Typically, absolute values of $t_i$ or $t_i^*$ that exceed 2 or 3 are considered evidence of a possible outlier with significance levels of .05 or .01, respectively. In other words, approximately 5% of “good” data will be declared “bad” data, because under normal theory $t_i$ are distributed as $t(n - k - 2)$.

When multiple outliers are being tested for statistical significance, individual significance levels are no longer valid. Cook & Weisberg (1982a) suggested using the Bonferroni inequality to obtain the $p$ values for outliers. Outliers may or may not be influential data points, but a substantial difference between $t_i$ and $t_i^*$ suggests the presence of an influential point. In any case, the presence of an outlier should at least alert the analyst of the possible uniqueness of the data point.

**Volume Measures**

The preceding measures indicate the unusualness of an observed point either with respect to the regressor values (along the horizontal axes), or with respect to the vertical deviations from the regression model. A given data point might not be unusual in either of these respects, but might be significant in their particular combinations. For example, a data point might have a moderately extreme value of both its leverage and residual. Thus, there is a need for influence measures that combine these into an overall indication of influence. Measures based on volume or distance use this idea.

Append the $y$ vector onto the design matrix $x$, and let $x^*$ be the resulting $n \times (k + 2)$ matrix, $x^* = (x, y)$. The determinant of $(x^*)^T x^*$, denoted by $||x^*||$ is proportional to the square root of the ellipsoidal volume of the $n$ vectors in the $(k + 2)$-dimensional Euclidean space. With the $i$th data point deleted, the volume spanned by the $(n - 1)$ vectors is obtained.

The Andrew-Pregibon $(A)$ measure (Cook & Weisberg, 1982a) is

$$A_i = \frac{||x_i^*||}{||x_0^*||},$$

If this ratio is close to 1, the $i$th data point is similar to the rest of the observations and is therefore not an influential point. If $A_i$ is much larger than 1, then the $i$th data point may be influential.

Volume measures are intuitively appealing. In two dimensions, if the space enclosed by the scatter of points is very different without a point that
is suspected to be an outlier or influential point, then that point may indeed be an outlier or an influential point. Thus, volume ratios like $A_i$ alert the analyst to the presence of possible data points that require further study. A value of $A_i$ greater than 2 suggests a problematic data point. More importantly, all $A_i$ values should be examined and if a particular value stands apart from the rest, then that particular observation requires special scrutiny.

It can be shown that $A_i = 1 - p_i^*$, where $p_i^*$ is the $i$th diagonal element of $P^*$; $P^* = x^* [(x^*)'x]^{-1}(x^*)'$, and $x^*$ is the extended matrix. A similar but slightly different ratio is called the covariance ratio ($C$) and is defined by

$$C_i = A_i \frac{s^2}{s_{ii}^*} ,$$

where $s^2$ and $s_{ii}^*$ are defined above. Belsley et al. (1980) suggested informal tests based on $C$ for detecting influential observations. In particular, they suggest that if the absolute value $|C_i - 1|$ is greater than $3(k + 1)/n + 1$, then the corresponding point should be considered as a possible candidate for being influential and/or an outlier.

Measures Based on Distance

If an observation is influential, the entire estimated vector $b$ is affected and each component is influenced by different amounts. If $b$ and $b_{i0}$ are the vectors of regression estimates, with and without the $i$th data point, then Cook's distance ($D_i$) is a standardized measure of the distance between the estimated vectors $b$ and $b_{i0}$, and is given by

$$D_i = \frac{[b - b_{i0}]'(x^*)'[(x^*)'x]^{-1}(x^*) b - b_{i0})}{(k + 1)s^2} ,$$

which also is equal to

$$D_i = \frac{t_i^2 \cdot h_i}{1 - h_i} .$$

A large value of $D_i$ relative to the $C_i$ values of other data points usually indicates that the $i$th data point is an influential point. There are other versions of $C_i$, such as that proposed by Welsch & Kuh (1977), and these all are variations of the well-known Mahalanobis distance (Belsley, Kuh, & Welsch, 1980).

The Welsch-Kuh (1977) distance ($W_i$), also called DFFITS by Belsley et al. (1980), is a measure of the effect of the $i$th observation on the $i$th predicted point and is measured by scaling the change in prediction at $x_i$ when the $i$th observation is deleted. Thus,

$$W_i = \frac{|x_i'(b - b_{i0})|}{s_{i0}\sqrt{h_i}} ,$$

which also is equal to

$$W_i = t_i \sqrt{\frac{h_i}{1 - h_i}} .$$

$W_i$ is similar to a $t$ statistic, and a value greater than $[(k + 1)/n]^{1/2}$ has been suggested in the literature as a warning sign to examine the data point carefully.

Selection of Diagnostics

The regression diagnostics literature contains 25 or more diagnostic indexes. Many of these measures convey similar information, and no single measure is fully informative or definitive in diagnostic assessments. Three to five measures that are easily available from common statistical software programs should be sufficient for a comprehensive data analysis in most cases, particularly if they are supplemented by graphical tools. Five useful diagnostics include the externally studentized residuals, diagonal elements of $H$, $D_i$, $W_i$, and $C_i$.

These five statistics were selected for the following reasons. The studentized residual yields information about outliers. The diagonal elements $h_i$ of $H$ provide information about the influence independent of the value of the response variable. $D_i$ gives the change in the parameter estimates. $C_i$ provides an indication of the aloofness (i.e., a distinct separation of the point from the rest of the data points) that may be due
to its influence or because it is an outlier. Finally, $W_i$ provides a numerical measure of the contribution of a data point to the overall fit of the model. These five measures cover the important aspects of regression diagnostics and the various vital statistics used for studying the appropriateness of a model. When used in conjunction with each other, these measures should allow successful detection of influential points in most situations.

**Graphical Diagnostics**

The discussion above centered on influential observations assuming that the number of regressors $k$ is fixed. However, the impact of a regressor variable on the estimates of other regression coefficients also may be studied. Such influences are called partial influences, and various diagnostics are available to study them. Partial leverage (or added variable plots) and augmented partial residual plots are examples of such diagnostics, in addition to the usual residual plots (Atkinson, 1985; Chatterjee & Ali, 1988; Cook, 1986; Cook & Weisberg, 1982a, 1982b; Daniel & Wood, 1980; Mosteller & Tukey, 1977, discussed graphical influence measures from the perspective of differential geometry). Partial leverage plots are briefly discussed here; for further discussions refer to the papers cited above.

A partial leverage plot for the $m$th explanatory variable is a plot of two sets of residuals. The first set is obtained by regressing $y$ on $x_1, x_2, \ldots, x_{m-1}, x_{(m+1)} \ldots, x_k$; the second set of residuals is obtained by regressing $x_m$ on $x_1, x_2, \ldots, x_{(m-1)}, x_{(m+1)} \ldots, x_k$. This removes the linear effect of the other regressors from both $y$ and $x_m$. The slope of the regression line between the two sets of residuals is the same as the regression coefficient of $x_m$ when $y$ is regressed on $x_1, x_2, \ldots, x_k$, hence the name partial leverage (or added variable) plots. Many other properties of partial leverage plots can be found in Velleman & Welsch (1981).

Diagnostics for Colinearity

The discussion above implicitly has assumed that the regressor variables are more or less independent. However, colinearity is often present in practice. Estimation in the presence of multicolinearity leads to larger standard errors of the parameter estimates. Remedies for this problem include ridge regression and ridge estimation (Delaney & Chatterjee, 1986).

If multicolinearity is present, the effect of a regressor variable on the variability of the response variable cannot be isolated from the effects of other explanatory variables. Although this does not prevent the use of the regression model for prediction purposes, it is a serious drawback when the main objective of the model is to understand and explain relationships, as is often true in behavioral research. To detect multicolinearity, examination of the correlation matrix of independent variables is often the first step. However, if an independent variable is a linear combination of several explanatory variables, then the correlation matrix is not sufficient. In this case, the variance inflation factors ($VIF$s) are recommended. For information on practical methods for detecting multicolinearity, see Mansfield & Helms (1982).

The precision of a least squares estimate is measured by its variance, which is proportional to $\sigma^2$, the variance of errors. The constant of proportionality is called $V_r$ for $\beta_r$, corresponding to the least squares estimate $b_r$ of each parameter $\beta_r$, given by

$$V_r = \frac{1}{1 - R^2_r}$$

where $R^2_r$ is the square of the multiple correlation coefficient from the regression of the $j$th explanatory variable on all other explanatory
variables of the regression equation. The denominator \((1 - R_j^2)\) is sometimes called "tolerance" (Stewart, 1987). If \(R_j^2\) is close to 1, indicating a strong relationship of variable \(j\) to the explanatory variables, then \(V_j\) will be large and tolerance will be low. On the other hand, if \(R_j^2\) is close to 0, \(V_j\) will be near 1. The literature suggests that \(V_j\) in excess of 5 to 10 is an indication that multicollinearity may be a problem.

A more sensitive measure for colinearity, called the condition number, has been studied by Stewart (1987). The condition number is the ratio of the largest to the smallest eigenvalue of the inverse of the sum of squares matrix \(x'x\). Although \(V_j's\) and condition numbers are related, the latter are easier to work with (from a theoretical and computational standpoint) for studying colinearity arising from a single data point or a group of data points and for studying the numerical precision of the parameter estimates in least squares estimation. However, the \(V_j's\) are more specific to covariates, local in nature, and more useful for the practitioner. Condition number, on the other hand, describes the overall colinearity of the regressors and, therefore, is global. For a further discussion of \(V_j\), condition number, and their impact on least squares estimates, see Stewart (1987), Belsley et al. (1980), and Chatterjee & Price (1973).

Because researchers in behavioral sciences often use scaled data, it is important to know the effect of scale changes on the computed diagnostics. None of the diagnostics presented here that deals with outliers, influential points, and measures of colinearity is invariant to nonlinear transformations. Statistics based on eigenvalues of \(x'x\), such as the condition number, are not invariant to linear changes in scale. Casella (1983) employed concepts of leverage to decide whether or not to include an intercept in the model. Other diagnostics based on measures of regression, residuals, volume, and distance for computing influence and leverage diagnostics are invariant to linear scale changes. Diagnostics should be interpreted within the context of the particular model being investigated, including the scales used.

**Regression Diagnostics in Statistical Software**

Most statistical software packages such as SPSS (SPSS, Inc., 1988), SAS (SAS Institute, Inc., 1982), BMDP (Dixon, 1984), and MINITAB (Ryan et al., 1991) include a variety of regression diagnostics. In SPSS, the REGRESSION command has options that produce various forms of residuals, elements of the prediction matrix (leverage), and various partial residual plots. In SAS, the PROC REG command has the subcommands COLLIN, \(V_j\), INFLUENCE, and RESIDUAL that produce eigenvalues, condition number, \(V_j's\), \(D_j\), the prediction matrix, \(C_j\), \(W_i\) (called DFFITS), and various other residuals. Cook & Weisberg (1982b) listed the options and commands that SAS, SPSS, and BMDP offer and a set of MINITAB instructions for computing various diagnostics. For users of microcomputers, SYSTAT (Wilkinson, 1992) provides a fairly extensive set of regression diagnostics including OLS, studentized (internal) and partial residuals, leverage, \(C_j\), and the standard errors of prediction. SYSTAT also has the capacity to plot these against the observation index (case) number, and the estimated and actual \(y\) values. The regression output includes the case numbers of extreme diagnostic values.

**Data Analysis Using Diagnostics**

The data consisted of a random sample of 24 patients who filled out a questionnaire. The variables were as follows:

1. \(Y\), perceived satisfaction level,
2. \(X_1\), patient’s age in years,
3. \(X_2\), severity of illness index, and
4. \(X_3\), level of anxiety felt.

\(X_1\) and \(X_2\) were obtained from patient surveys.
and other records. Information for $X_i$ was provided by the resident physician. With the exception of $X_i$, these variables were measured on imprecise subjective rating scales and are typical of the variables commonly used by behavioral and social scientists. The data are provided in Table 1.

### Table 1
Patient Satisfaction Data for Four Variables

<table>
<thead>
<tr>
<th>Person</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43</td>
<td>53</td>
<td>2.4</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>48</td>
<td>2.4</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>29</td>
<td>50</td>
<td>2.1</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>52</td>
<td>62</td>
<td>2.9</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>48</td>
<td>48</td>
<td>2.4</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>42</td>
<td>50</td>
<td>2.2</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>49</td>
<td>54</td>
<td>2.9</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>28</td>
<td>43</td>
<td>1.8</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>41</td>
<td>44</td>
<td>1.8</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>63</td>
<td>65</td>
<td>1.7</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>40</td>
<td>48</td>
<td>2.2</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>36</td>
<td>46</td>
<td>2.3</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>50</td>
<td>51</td>
<td>2.3</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>38</td>
<td>55</td>
<td>2.2</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>34</td>
<td>51</td>
<td>2.3</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>53</td>
<td>54</td>
<td>2.2</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>36</td>
<td>49</td>
<td>2.3</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>33</td>
<td>56</td>
<td>2.5</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>29</td>
<td>46</td>
<td>1.9</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>33</td>
<td>49</td>
<td>2.1</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>55</td>
<td>51</td>
<td>2.4</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>29</td>
<td>52</td>
<td>2.3</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>44</td>
<td>58</td>
<td>2.9</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>43</td>
<td>50</td>
<td>2.3</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>

The proposed model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad (15)$$

and the estimated regression equation is given by

$$\hat{y}_i = 155.11 - 1.15x_{i1} + .27x_{i2} - 15.05x_{i3} \quad (16)$$

(6.88)  (-4.1)  (-.5)  (-2.09)

(the calculated t values are given in parentheses in Equation 16, and in other equations below).

For this equation, $R_{adj}^2 = .634$, $F = 14.38$, $p < .001$. For Equation 16, $X_i$ is not significant. Equation 16 is referred to as Model 1 (see Table 2).

Examination of the residuals in Table 2 does not provide evidence of any possible outliers. However, from examination of $D_i$, $h_i$, $W_i$, and $C_i$, Person 10 appears to be an influential point. Figures 2a, 2b, and 2c provide a plot of $h_i$, $W_i$, and $C_i$ for each person. Figure 2 and Table 2 show that Person 10 is a highly influential point. This point deserves careful scrutiny and often raises interesting questions, such as: Have the data been recorded or transmitted incorrectly? Is the influence due to additional variables that have...
<table>
<thead>
<tr>
<th>Model and Person</th>
<th>Residuals</th>
<th>Distance</th>
<th>Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t_i$</td>
<td>$t^a_i$</td>
<td>$D_i$</td>
</tr>
<tr>
<td><strong>Model 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>11.567</td>
<td>1.170</td>
<td>1.182</td>
</tr>
<tr>
<td>2</td>
<td>16.184</td>
<td>1.707</td>
<td>1.800</td>
</tr>
<tr>
<td>3</td>
<td>.209</td>
<td>.022</td>
<td>.022</td>
</tr>
<tr>
<td>4</td>
<td>-9.172</td>
<td>-1.073</td>
<td>-1.077</td>
</tr>
<tr>
<td>5</td>
<td>-.493</td>
<td>-.051</td>
<td>-.051</td>
</tr>
<tr>
<td>6</td>
<td>-14.398</td>
<td>-1.458</td>
<td>-.051</td>
</tr>
<tr>
<td>7</td>
<td>-4.768</td>
<td>-.540</td>
<td>-.530</td>
</tr>
<tr>
<td>8</td>
<td>4.659</td>
<td>.513</td>
<td>.503</td>
</tr>
<tr>
<td>9</td>
<td>.817</td>
<td>.091</td>
<td>.089</td>
</tr>
<tr>
<td>10</td>
<td>3.178</td>
<td>.665</td>
<td>.655</td>
</tr>
<tr>
<td>11</td>
<td>2.771</td>
<td>.283</td>
<td>.276</td>
</tr>
<tr>
<td>12</td>
<td>-9.845</td>
<td>-1.025</td>
<td>-1.026</td>
</tr>
<tr>
<td>13</td>
<td>-1.462</td>
<td>-.153</td>
<td>-.150</td>
</tr>
<tr>
<td>14</td>
<td>-16.629</td>
<td>-1.740</td>
<td>-1.840</td>
</tr>
<tr>
<td>15</td>
<td>-16.785</td>
<td>-1.171</td>
<td>-.813</td>
</tr>
<tr>
<td>16</td>
<td>10.279</td>
<td>1.085</td>
<td>1.091</td>
</tr>
<tr>
<td>17</td>
<td>-4.549</td>
<td>-.467</td>
<td>-.458</td>
</tr>
<tr>
<td>18</td>
<td>14.430</td>
<td>1.595</td>
<td>1.664</td>
</tr>
<tr>
<td>19</td>
<td>7.120</td>
<td>.758</td>
<td>.750</td>
</tr>
<tr>
<td>20</td>
<td>-12.480</td>
<td>-1.280</td>
<td>-1.303</td>
</tr>
<tr>
<td>21</td>
<td>6.769</td>
<td>.758</td>
<td>.749</td>
</tr>
<tr>
<td>22</td>
<td>3.759</td>
<td>.404</td>
<td>.395</td>
</tr>
<tr>
<td>23</td>
<td>6.586</td>
<td>.740</td>
<td>.731</td>
</tr>
<tr>
<td>24</td>
<td>2.252</td>
<td>.229</td>
<td>.223</td>
</tr>
<tr>
<td><strong>Model 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>12.132</td>
<td>1.214</td>
<td>1.231</td>
</tr>
<tr>
<td>2</td>
<td>13.858</td>
<td>1.350</td>
<td>1.614</td>
</tr>
<tr>
<td>3</td>
<td>.606</td>
<td>.063</td>
<td>.062</td>
</tr>
<tr>
<td>4</td>
<td>-7.675</td>
<td>-.917</td>
<td>-.913</td>
</tr>
<tr>
<td>5</td>
<td>-.177</td>
<td>-.190</td>
<td>-.185</td>
</tr>
<tr>
<td>6</td>
<td>-13.799</td>
<td>-1.383</td>
<td>-1.420</td>
</tr>
<tr>
<td>7</td>
<td>-6.634</td>
<td>-.781</td>
<td>-.773</td>
</tr>
<tr>
<td>8</td>
<td>4.150</td>
<td>.452</td>
<td>.442</td>
</tr>
<tr>
<td>9</td>
<td>1.550</td>
<td>.172</td>
<td>.168</td>
</tr>
<tr>
<td>10</td>
<td>2.449</td>
<td>.247</td>
<td>.241</td>
</tr>
<tr>
<td>11</td>
<td>-11.863</td>
<td>-1.283</td>
<td>-1.307</td>
</tr>
<tr>
<td>12</td>
<td>-.589</td>
<td>-.062</td>
<td>-.060</td>
</tr>
<tr>
<td>13</td>
<td>-14.310</td>
<td>-1.585</td>
<td>-1.656</td>
</tr>
<tr>
<td>14</td>
<td>-16.954</td>
<td>-1.710</td>
<td>-1.810</td>
</tr>
<tr>
<td>15</td>
<td>13.179</td>
<td>1.545</td>
<td>1.608</td>
</tr>
<tr>
<td>16</td>
<td>-3.449</td>
<td>-.354</td>
<td>-.346</td>
</tr>
<tr>
<td>17</td>
<td>14.888</td>
<td>1.627</td>
<td>1.707</td>
</tr>
<tr>
<td>18</td>
<td>7.220</td>
<td>.758</td>
<td>.749</td>
</tr>
<tr>
<td>19</td>
<td>-12.219</td>
<td>-1.237</td>
<td>-1.255</td>
</tr>
<tr>
<td>20</td>
<td>7.324</td>
<td>.812</td>
<td>.804</td>
</tr>
<tr>
<td>21</td>
<td>3.660</td>
<td>.388</td>
<td>.379</td>
</tr>
<tr>
<td>22</td>
<td>5.978</td>
<td>.666</td>
<td>.656</td>
</tr>
<tr>
<td>23</td>
<td>2.272</td>
<td>.228</td>
<td>.222</td>
</tr>
</tbody>
</table>
not been included in the model, or due to interaction among the existing variables? Is the point simply an aberration? These questions may lead the researcher to modify the model, search for other explanatory variables, gather additional data, or drop the influential point from the analysis. To demonstrate the effect of this observation on the current model, it was excluded from the dataset.

The regression for the reduced dataset without Person 10 is referred to as Model 2. The estimated regression equation is given by

\[ \hat{y} = 162.88 - 1.21x_{11} - 0.67x_{12} - 8.61x_{13}, \]  

which resulted in \( R_{adj}^2 = .621 \), with \( F = 13.01, p < .001 \). In Model 2, both \( X_2 \) and \( X_3 \) are not significant.

There is a large difference between the two models even though an explanatory variable which was highly significant in Model 1 became insignificant in Model 2. Moreover, removal of Person 10 revealed possible multicolinearity between \( X_2 \) and \( X_3 \). The \( V_i^2 \) values in Model 1 were 1.6, 1.8, and 1.2 for variables \( X_1, X_2, \) and \( X_3 \), respectively. The highest correlation was between \( X_1 \) and \( X_2 \) (\( r = .61 \)). On the other hand, the \( V_i^2 \)'s for Model 2 were 1.4, 2.8, and 2.9. The correlation between \( X_1 \) and \( X_2 \) was approximately .80, indicating strong collinearity. Thus, the removal of a single data point can reveal considerable multicollinearity. Extended discussions of multicollinearity from insertion and deletion of a small set of data points can be found in Chatterjee & Hadi (1988).

The diagnostics after deleting Person 10 also are given in Table 2, and plots of \( h_i, W_i, \) and \( C_i \) by person for this reduced dataset are given in Figure 3; these results do not reveal any influential data point.

If \( X_1 \) or \( X_3 \) is dropped, the regression equations are

\[ \hat{y} = 166.59 - 1.26x_{11} - 1.09x_{12}, \]  

with \( R_{adj}^2 = .631 \)

and

\[ \hat{y} = 147.08 - 1.243x_{11} - 8.256x_{13}, \]  

with \( R_{adj}^2 = .627 \); the diagnostic values showed no influential points.

**Partial Residual Plot**

The partial residual plot for \( X_2 \) was constructed by regressing \( Y \) on \( X_1 \) and \( X_2 \) and retrie-
ing the residuals (called Res1), and by regressing 
X3 on the regressors X1 and X2 and retrieving 
the residuals (called Res2). Figure 4 is the partial 
leverage plot for X3 with Res1 (vertical axis) 
against Res2 (horizontal axis). Note the influen-
tial nature of Person 10. Without Person 10, the 
statistical significance of X3 becomes doubtful. 
Similar plots for other regressor variables yielded 
identical conclusions. In general, it is advisable 
to look at the partial regression plots for all the 
explanatory variables to detect possible misspeci-
fication in functional form, heteroscedasticity, 
and influential points and/or outliers.

Figure 4
A Partial Leverage Plot

There are two additional possibilities, other 
than gross errors, that may explain the data for 
the deleted Person 10. One is the possibility of 
interaction between regressor variables, and the 
other is the relevance of other unmeasured ex-
planatory variables. These possibilities must be 
investigated before final conclusions can be 
reached. Thus, a researcher faced with the 
analysis of actual data may proceed in two ways: 
In the absence of any relevant information on 
errors in variables, the usual regression diagnost-
cs should be used to detect any influential or 
outlier points and proceed accordingly. On the 
other hand, if there is information available on 
the error structure of the regressors, the re-
searcher can use the structural equations ap-
proach or study the sensitivity of the regression 
coefficients by any three of the methods men-
tioned earlier.

Discussion

Do regression diagnostics help uncover 
substantive issues? For the full dataset, the 
variability in Y was explained by X1 and X2, but 
X3 was not statistically significant. However, 
when the data for Person 10 were dropped from 
the analysis, both X1 and X2 were jointly not 
significant. These results suggest that X1 and 
either X2 or X3 should be used to explain the 
variability in Y. This parsimonious description of 
the data was facilitated through the use of regres-
sion diagnostics.

Such findings are not limited to small data-
sets (Belsley et al., 1980). Huber (1981) in-
vestigated robustness properties of least squares 
estimates and concluded that 1 to 5% of the 
number of observations that are influential can 
affect the results of a regression analysis. The 
percentage of influential points that significantly 
distort or bias the parameter estimates of a model 
is called the "breakdown point of an estimator" 
(Rousseeuw & LeRoy, 1987). Theoretical investi-
gations have shown that the breakdown point of 
least squares estimates approaches 0 as the 
number of observations becomes large. Conse-
quently, even a small percentage of influential 
points in large datasets can affect the results, em-
phasizing the need to examine various diagnostics 
irrespective of the sample size.

Robust statistics (Huber, 1981) allow the 
analyst to select a weight between 0 and 1 for each 
data point, which enables the observation to ex-
ert influence on the results proportional to its 
weight. A weight of 0 implies deletion, and a 
weight of 1 leaves the observation in the estima-
tion sample. The researcher who is acquainted 
with both the problem and the dataset must 
decide how to treat an influential observation. 
Regression diagnostics can only point to the 
presence of influential observations.

Although commonly available regression 
diagnostics are adequate in the context of linear 
regression models, there are some situations in
which using regression diagnostics are inadequate. The paper by Huber (1983) and the discussions following by several researchers and the subsequent rejoinder by Huber provide a helpful guide to the applied researcher. A study of the impact of the usual regression diagnostics when the assumptions of the regression model are violated (particularly the assumption of the independence of errors) was studied by Polasek (1984). The behavioral researcher should be aware of situations in which the usual diagnostics may be inadequate.

Multiple Influential Points

Multiple influential points are common in practical datasets. Provided that the influential points do not mask each other, the methods described above are adequate for their detection. Masking occurs when one influential point hides the presence of others. Figure 5 shows a simple case of masking. The least squares line is given by AB. Points 3 and 4 are individually influential, but neither Point 1 or 2 appears to be influential by itself. According to some researchers, masking is not a problem in practice (see Chatterjee & Hadi, 1986).

Figure 5

Example of Individually Influential Data Points (3 and 4) and Masked Jointly Influential Points (1 and 2)

The data for Figure 5 (reproduced in the second and third columns of Table 3) are used to illustrate the masking phenomenon. The OLS regression of these points is given by

$$\hat{y} = 5.28 + 1.28x$$

and with Observation 2 deleted (but including Observation 1),

$$\hat{y} = 4.71 + 1.48x$$

However, if both Observations 1 and 2 are deleted,

$$\hat{y} = 3.72 + 1.87x$$

The individual regression diagnostics for Observations 1 and 2 were small, but the absence of these points changed the slope of the regression estimate from 1.28 to 1.87, and changed the intercept from 5.28 to 3.72.

When masking is present, jointly influential points can be detected by computing the diagnostics after removing pairs, triplets, and so forth of points from the dataset. However, for any reasonable dataset, the number of computations become unmanageable. There are several strate-
gies to deal with this problem.

The first approach uses the off-diagonal elements of \( H \). The off-diagonal elements of \( H, h_{ij} \), play a significant role in the determination of joint influence. For example, a high value of \( h_{ij} \) may indicate the presence of the joint influence of the \( i \)th and the \( j \)th observations. Gray & Ling (1984) detected influential subsets in regression using a clustering algorithm. They provided a graphical method for detecting jointly influential points. For a more complete discussion of the use of off-diagonal elements of \( H \) for detecting joint influence, see the discussion following Gray & Ling (1984).

The second strategy advocated by Rousseeuw (1984) and Rousseeuw & Leroy (1987) uses the least median squares (LMS) approach. With LMS, the estimates are obtained by minimizing the median squared residuals and are robust in the presence of multiple outlier points. The LMS estimates can be thought of as determining a median or modal plane (model) and are not influenced by multiple outliers. Rousseeuw and Leroy discussed the properties of LMS estimates and provided an algorithm for determining the LMS estimates and subsequent detection of outliers. Rousseeuw & Van Zomeren (1990) provided an alternative approach, based on the Mahalanobis distance, for unmasking multivariate outliers and leverage points.

A third strategy considers the data as \( n \) points in a \((k + 1)\)-dimensional space. The vertices of the convex hull (smallest convex set) of these points are proposed as candidates for testing for joint influence. Figure 6 presents the outermost convex hull of a set of points in two-dimensional space. The vertices of the outermost layer (peel) are denoted by Points 1 through 8. This layer of points may be removed in order to continue examining vertices of inner hulls. This approach, which reduces the exhaustive search for multiple influential points, was proposed by Chatterjee & Chatterjee (1990). It is still in an experimental stage, as is the suggestion to examine each vertex together with the points that are in the vicinity of that vertex. The convex hull approach was used successfully by Chatterjee, Jamieson, & Wiseman (1991) for detecting multiple influential points in factor analysis. Peeling of convex hulls for ordering multivariate data can be found in Green (1984), and other uses of the vertices of convex hulls as extreme points are discussed in Barnett (1976).

Diagnostics for Multivariate Regression and Other Models

Multivariate Diagnostics

There appears to be very little work done on multivariate diagnostics, other than the work of Hossain & Naik (1989). Residual-based, volume-based, and distance-based diagnostics for the multivariate regression model are discussed briefly. There appears to be no commonly available software to implement these, but the diagnostics described below can be computed using the PROC MATRIX command in SAS (SAS Institute, Inc., 1982).

Consider the multivariate regression model

\[
Y = XB + E,
\]

where \( Y \) is a \( n \times p \) response matrix of \( n \) observations of \( p \) response variables, \( X \) is a \( n \times (k + 1) \) matrix of regressors, \( B \) is a \((k + 1) \times p \) matrix of parameter values, and \( E \) is a \( n \times p \) matrix of residuals.
The columns of $\mathbf{B}$ are denoted by $b_j$. $\mathbf{X}$ is assumed to be nonstochastic and of full rank, [i.e., $\text{rank}(\mathbf{X}) = k + 1$]. Rows $e_i$ of $\mathbf{E}$ are assumed to be independent and normally distributed, with a $p \times 1$ mean vector of $0$s and a $p \times p$ covariance matrix $\Sigma$. The least squares estimates of $b_j$ are given by

$$\hat{b}_j = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

$$j = 1, 2, \ldots, p,$$

which is the same as equation-by-equation least squares estimation.

First, note that the diagonal elements $h_i$ of $\mathbf{H}$, the prediction matrix, play the same role as in multiple regression. Large values indicate that at least one component of the $i$th data point may be an influential point.

The residual matrix is given by $\mathbf{E} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}$. An estimate of $\Sigma$ is

$$\hat{\Sigma} = \frac{\mathbf{E}'\mathbf{E}}{n - k - 1}.$$  

(26)

Let

$$\hat{\Sigma}_i = \frac{\mathbf{E}_{(i)}\mathbf{E}_{(i)}}{n - k - 1},$$

(27)

where $\mathbf{E}_{(i)} = \mathbf{Y}_{(i)} - \mathbf{X}_{(i)}\hat{\mathbf{b}}_{(i)}$ is obtained by dropping the $i$th observation. Define

$$\tau_i^2 = \frac{1}{(1 - h_i)} e_i' \hat{\Sigma}^{-1} e_i,$$

(28)

and

$$T_i^2 = \frac{1}{(1 - h_i)} e_i' \hat{\Sigma}_i^{-1} e_i,$$

(29)

where $\tau_i^2$ and $T_i^2$ are the multivariate analogues of $t_i$ and $t_1^2$ discussed earlier. These statistics are approximately distributed as Hotelling's $T^2$ distribution and can be tested for an outlier by comparing their values to the upper percentage point of an $F$ distribution with appropriate degrees of freedom.

A multivariate analogue of $D_i$ is given by

$$D_i^\circ = \frac{\tau_i^2}{k + 1} \frac{h_i}{1 - h_i}, \quad i = 1, 2, \ldots, n.$$  

(30)

Influence of the $i$th observation on the estimate of $\hat{\mathbf{B}}$ is indicated by large values of $D_i^\circ$. $W_i$ can be modified accordingly as

$$W_i^\circ = \frac{h_i}{1 - h_i} T_i^2,$$

$$i = 1, 2, \ldots, n,$$

and a cutoff value is given by

$$k + 1 \frac{p(n - k - 2)}{n - k - p - 1} F_{\alpha, p, n - k - p - 1},$$

(32)

where $F_{\alpha, p, n - k - p - 1}$ is an upper percentage point of the $F$ distribution with $p$ and $(n - k - p - 1)$ degrees of freedom.

$C_i^\circ$ is the corresponding multivariate analogue of the scalar variances and is given by

$$C_i^\circ = \left( \frac{1}{1 - h_i} \right)^p \left( \frac{||\hat{\Sigma}_i||}{||\hat{\Sigma}||} \right)^{1/1}.$$  

(33)

As before, very high or very low values of $C_i^\circ$ are considered to be significant. Hossain & Naik (1989) discussed other multivariate diagnostic measures for regression.

Nonlinear Models

Many nonlinear models and generalized linear models—including logistic regression and survival models—(McCullagh & Nelder, 1984) have become popular in the social sciences due to the enhanced capability they offer in building explanatory models. Nonlinear regression parameters can be very sensitive to values that are extreme in the response space or in the explanatory variables. One particular difficulty with nonlinear regression is that the diagnostics cannot be obtained in closed analytic form as in linear regression, and specific diagnostic routines must be built into each nonlinear regression program. Diagnostics for logistic regression and their applications to other nonlinear models have been...
described by Pregibon (1981) and Jorgenson (1983). Unfortunately, no suitable computer package exists that computes and provides nonlinear regression diagnostics on a routine basis.

Other Multivariate Methods

Influential data points also can exert considerable influence on the results of principal components and factor analysis models. The estimates of eigenvalues, eigenvectors, factor loadings, and factor scores can be altered dramatically by the presence of a small set of influential observations. A statistic similar to \( A \), was proposed by Chatterjee et al. (1991) that detects influential observations in principal components and factor analysis models. Computational requirements also are discussed by these authors, but commercial computer packages for detecting influential data points in principal and factor analysis models are not yet available.

Summary and Conclusions

Regression models need to be examined beyond the usual summary statistics and tests of significance. This is especially important when behavioral variables are involved, because of the inherent difficulties and errors in the measurement of these variables. Five regression diagnostics were recommended for a sound analysis of data: (1) \( h_i \), the diagonal elements of the prediction matrix; (2) \( t_i^* \), the externally studentized residual; (3) \( D_n \), Cook's distance; (4) \( W_n \), the Welsch-Kuh distance; and (5) \( C_n \), the covariance ratio. These numeric measures should be supplemented with graphical tools, such as a variety of residual plots.

Use of regression diagnostics often reveals data points whose presence will affect the estimates and significance of the parameters. As low as 1% of influential data points may render one or more variables significant or insignificant, may induce or remove multicollinearity, and may change estimates of parameter vectors and other statistics in an unpredictable way.

Applications of regression diagnostics in other multivariate models such as cluster analysis, and causal models such as path analysis and LISREL models, have not yet appeared in the literature. Developments of both theory and computer packages to detect outliers in nonlinear regression, weighted least squares, two-stage least squares, and other multivariate models also should be useful.

The interplay between measurement error, regression diagnostics, and cross-validation is an interesting one. Very little work has been done on the relationship between cross-validation and regression diagnostics and how to combine the procedures for useful data analysis. The relationship between measurement errors and regression diagnostics is easier to comprehend. Measurement errors, if arising due to an underlying causal mechanism, obviously can be modeled. However, when the underlying causal mechanisms giving rise to the measurement errors are unknown or difficult to model, then the measurement errors will be confounded with influential points and outliers. In that case, treatment of data by regression diagnostics will not address the problems arising from measurement errors.

Wellman & Gunst (1991) studied how influential observations affect linear measurement estimators. Their results indicate that the effects of influential observations are in a direction orthogonal to and along the fitted plane, rather than vertically and horizontally. They developed diagnostics patterned after least squares diagnostics, but these diagnostics follow the methods used for developing diagnostics for nonlinear regression and generalized linear models.

The techniques of regression diagnostics can detect influential points and what would happen if such points are removed from the analysis, but they do not answer many important questions. In particular, whether to keep or remove such data points cannot be answered in an abstract setting. Very often, bringing attention to a small set of unusual data points is a first step that can ultimately reveal much about the process under study. If the purpose of most data analysis is model building, regression diagnostics and...
various versions of cross-validation can be seen as different activities with a common purpose. Efron (1979) suggested that further research may lead to powerful combinations of cross-validation, the jackknife, and the bootstrap.

References


Acknowledgments

The authors appreciate the advice of two anonymous referees, which considerably improved the paper.

Author’s Address

Send requests for reprints or further information to Sangit Chatterjee, Northeastern University, 219 Hayden Hall, 360 Huntington Avenue, Boston MA 02115, U.S.A.