Chatterji, S & Yilmaz, M. (1992). A renewed regression diagnostics for behavioral research. Applied Psychological Measurement; 16, 209-227

Published quarterly and copyrighted by
it Inc., N657 Elliott Hall, University of
0344, U.S.A. All rights reserved. No part
d, stored in a retrieval system, or transmit-
lectronic, mechanical, photocopying, re-
prior written permission of the publisher.
bottom of the first page of an article in
owner's consent that copies of the ar-
ond that permitted by Sections 107 and
vided that the copier pay the stated per-
rance Center, Inc., 21 Congress St., Salem
xtend to other kinds of copying, such as
r advertising or promotional purposes,
r for resale, all of which require written
ting requests should be addressed to *Ap-*
57 Elliott Hall, University of Minnesota,
55455-0344 U.S.A.; phone 612-625-0862.
nd SUBMISSION OF MANUSCRIPTS:
of this issue.

A Review of Regression Diagnostics for Behavioral Research

Sangit Chatterjee and Mustafa Yilmaz
Northeastern University

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 nonrobustness of the least squares estimates in the presence of influential points is emphasized. Diagnostics for multiple influential points, multivariate regression, multicolinearity, 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-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 3, September 1992, pp. 209-227
© Copyright 1992 Applied Psychological Measurement Inc.
0146-6216/92/030209-19$2.20

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,

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \quad , \tag{1}$$

or $\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ in matrix notation, where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ 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 ith observation, $x_{i1}, x_{i2}, \ldots, x_{ik}$ are the observed values of k regressor (independent or predictor) variables $X_1, X_2, \ldots, X_k$. $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^{\mathrm{T}}$ is a $(k + 1) \times 1$ column vector of unknown constants, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is a $n \times 1$ column vector of random errors, and $\mathbf{x}$ 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 $b = (b_0, \ldots, b_k)^{\mathrm{T}}$, given by $b = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{y}$, which is obtained by minimizing $\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}$ with respect to $\boldsymbol{\beta}$. It is assumed that the random errors ε_i are independent with mean 0 and variance σ^2 (homoscedasticity). If the x_{ij}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 $b_1, b_2, \ldots, b_k$; the estimate of the standard error of the regression equation σ^2 is given by

$$s^2 = \frac{(\mathbf{y} - \mathbf{b})^{\mathrm{t}}(\mathbf{y} - \mathbf{b})}{n - k - 1} \tag{2}$$

and the measure of fit by

$$R^2 = (\hat{\mathbf{y}}\hat{\mathbf{y}}^{\mathrm{t}} - n\bar{y}^2)/\hat{\mathbf{y}}^{\mathrm{t}}\mathbf{y} - n\bar{y}^2 \quad , \tag{3}$$

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

$$R^2_{adj} = 1 - \left(\frac{n - 1}{n - k}\right)(1 - R^2) \quad . \tag{4}$$

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 β_js using generalized least squares estimation, which minimizes $\boldsymbol{\varepsilon}^{\mathrm{t}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}$, where $\boldsymbol{\Sigma}$ is the $n \times n$ variance-covariance matrix of $\boldsymbol{\varepsilon}$. 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 of the y_i, given $x = \{x_{ij}\}$, are independent with constant variance σ^2, and regressors are independent random variables with distributions that do not depend on the parameters $\boldsymbol{\beta}$ or the constant variance σ^2. Even correlated errors present no additional problems if it can be assumed that the distribution of x does not depend on $\boldsymbol{\beta}$ or $\boldsymbol{\Sigma}$.

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) f information on measurement error models]. A noted above, the presence of measurement err leads to estimates of the regression coefficien that are nonoptimal (i.e., biased or inconsisten The effect of this nonoptimality can be inv tigated using (1) the asymptotic approach, (2) t perturbation approach, and (3) the simulati approach.

In the asymptotic approach, which is popul in econometrics, large sample biases a calculated analytically and limiting values a sought (Chatterjee & Hadi, 1986). In the pertu bation approach, the effects on the regressi coefficients of perturbing the regressors by sm amounts are studied analytically. This meth provides an upper bound on the relative erro in the estimated coefficients (see Stewart, 198 In the simulation approach, the structure of t errors in the regressor variables is simulated, a the effects are studied on the estimated regre sion coefficients (Chatterjee & Hadi, 1986 These three approaches require the researcher make assumptions about the form of the distrib tions, the moments of the distributions, and/o the bounds on the errors in the variables. Ref to Chatterjee & Hadi (1988) for details of all thr methods. Fuller (1987) and Heller (1987) provi ed specialized discussions for studying vario aspects of measurement errors on the paramet estimates, and Bollen (1989) discussed modelin in the presence of these errors.

In theory, "instrumental variables" may pr vide a possible solution to the problem measurement errors in the regression variable Instrumental variables correlate very highly wi the regressors but not with the measureme errors in the regressors or the random erro associated with the dependent variable. If su variables can be found, they can be used regressor variables; alternatively, regress variables could be modeled in terms of the i strumental variables. This last approach leads the estimation of simultaneous equations by t two-stage least squares approach mention above (Judge et al., 1980). Clearly, the ma

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.

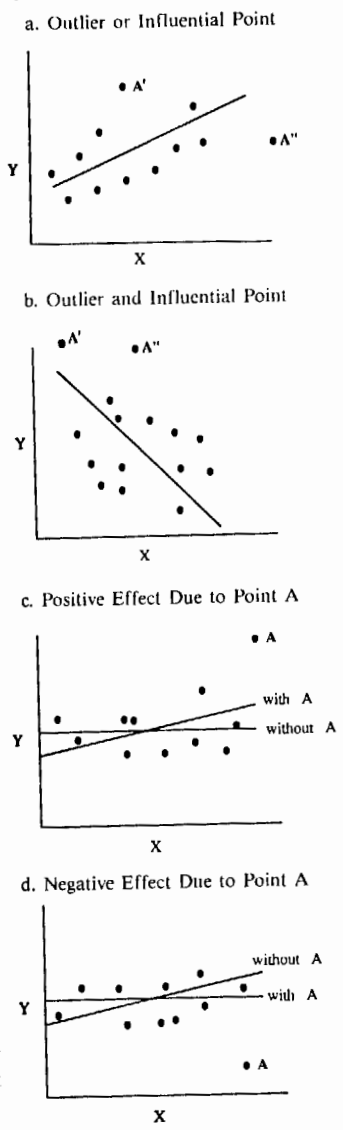
If influential points are present, serious errors may be made in interpreting the regression model. In extreme cases, it is possible to conclude that a hypothesized relationship exists when in fact it does not, or to conclude that no relationship exists when it does. In Figure 1c, point A is highly influential—the 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).

Figure 1
Examples of Outliers and Influential Points



Regressor-Based Measures

To study the impact of x and y on b, define a $n \times n$ prediction matrix, given by $\mathbf{H} = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T$. Denote the fitted values $\hat{\mathbf{y}}$, where $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, and $\mathbf{H}$ maps the observed values into the predicted values. The diagonal elements of $\mathbf{H}$ are denoted by h_i, and the off-diagonal elements are denoted by h_{ij}. From $h_i = x_i(\mathbf{x}^T\mathbf{x})^{-1}x_i^T$, it follows that the fitted value $\hat{y}_i$ depends on the ith 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_i 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 $h_i = \sum_{j=1}^{n} h_{ij}^2$, so that no entry h_{ij} can be larger in absolute value than $(h_i)^{1/2}$. In other words, the contribution of the jth observation on the ith 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 ith observation, an internally studentized residual is defined as

$$t_i = \frac{e_i}{s(1 - h_i)^{1/2}} \quad , \tag{5}$$

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

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - k - 1} \quad , \tag{6}$$

and the externally studentized residual is given by

$$t_i^* = \frac{e_i}{s_{(i)}(1 - h_i)^{1/2}} \quad , \tag{7}$$

where $s_{(i)}$ is the standard error of the regression equation obtained from deleting the ith observation. The internally studentized residual uses the standard error s_i in the scaling, whereas the externally studentized residual uses the standard error $s_{(i)}$ without the ith 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^*s 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 $\mathbf{x}$, and let $\mathbf{x}^*$ be the resulting $n \times (k + 2)$ matrix, $\mathbf{x}^* = (\mathbf{x}, \mathbf{y})$. The determinant of $(\mathbf{x}^*)^T\mathbf{x}^*$, denoted by $||(\mathbf{x}^*)^T\mathbf{x}^*||$ is proportional to the square root of the ellipsoidal volume of the n vectors in the $(k + 2)$-dimensional Euclidean space. With the ith data point deleted, the volume spanned by the $(n - 1)$ vectors is obtained.

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

$$A_i = \frac{||(\mathbf{x}^*)^T x^*||}{||(\mathbf{x}^*_{(i)})^T x^*_{(i)}||} \quad . \tag{8}$$

If this ratio is close to 1, the ith 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 ith 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 ith diagonal element of $\mathbf{P}^*$; $\mathbf{P}^* = \mathbf{x}^*[(\mathbf{x}^*)^T\mathbf{x}]^{-1}(\mathbf{x}^*)^T$, and $\mathbf{x}^*$ is the extended matrix. A similar but slightly different ratio is called the covariance ratio (C) and is defined by

$$C_i = A_i \frac{s^2}{s_{(i)}^2} \quad , \tag{9}$$

where s^2 and $s_{(i)}^2$ 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 $\mathbf{b}$ is affected and each component is influenced by different amounts. If $\mathbf{b}$ and $\mathbf{b}_{(i)}$ are the vectors of regression estimates, with and without the ith data point, then Cook's distance (D_i) is a standardized measure of the distance between the estimated vectors $\mathbf{b}$ and $\mathbf{b}_{(i)}$, and is given by

$$D_i = \frac{[\mathbf{b} - \mathbf{b}_{(i)}]^T(\mathbf{x}^T\mathbf{x})[\mathbf{b} - \mathbf{b}_{(i)}]}{(k + 1)s^2} \quad , \tag{10}$$

which also is equal to

$$D_i = \frac{t_i^2}{(k + 1)} \cdot \frac{h_i}{1 - h_i} \quad . \tag{11}$$

A large value of D_i relative to the C_i values of other data points usually indicates that the ith 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 ith observation on the ith predicted point and is measured by scaling the change in prediction at x_i when the ith observation is deleted. Thus,

$$W_i = \frac{|x_i^T(\mathbf{b} - \mathbf{b}_{(i)})|}{s_{(i)}\sqrt{h_i}} \quad , \tag{12}$$

which also is equal to

$$W_i = t_i^2 \sqrt{\frac{h_i}{1 - h_i}} \quad . \tag{13}$$

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 $\mathbf{H}$, 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 $\mathbf{H}$ 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 mth 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). A partial leverage plot for variable x_m may point out the functional form of its relationship with y (presence of curvilinearity), heteroscedasticity, and the presence of influential points or outliers. Partial leverage plots thus may be considered as a multivariate analogue of an ordinary bivariate scatterplot. For a demonstration of a practical use of partial leverage plots on two substantive problems in sociology, see Bollen & Jackman (1985).

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 (V_js) 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 σ^2, the variance of errors. The constant of proportionality is called V_j. There is a V_j corresponding to the least squares estimate b_j of each parameter β_j, given by

$$V_j = \frac{1}{1 - R_j^2} \quad , \tag{14}$$

where R_j^2 is the square of the multiple correlation coefficient from the regression of the jth 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 multicolinearity 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 $\mathbf{x}'\mathbf{x}$. Although V_js 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_js 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 the 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 $\mathbf{x}'\mathbf{x}$, such as the condition number, are not invariant to linear changes in scales nor are most of the procedures based on ridge regression. There is a great deal of discussion in the literature as to whether condition numbers should be calculated on untransformed data or mean-centered data. Belsley (1984; see also the following discussions and the rejoinder by Belsley) shed light on this controversial issue. Casella (1983) employed concepts of leverage to decide whether or not to include an intercept in the model. Other diagnostics based on measures of regressors, 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_js, D_i, the prediction matrix, C_i, 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_i, 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_2 was provided by the resident physician. With the exception of X_1, 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

Person	X_1	X_2	X_3	Y
1	43	53	2.4	67
2	29	48	2.4	89
3	29	50	2.1	77
4	52	62	2.9	26
5	45	48	2.4	54
6	42	50	2.2	46
7	49	54	2.9	36
8	28	43	1.8	89
9	41	44	1.8	70
10	63	65	1.7	43
11	40	48	2.2	66
12	36	46	2.3	57
13	50	51	2.3	48
14	38	55	2.2	47
15	34	51	2.3	51
16	53	54	2.2	57
17	36	49	2.3	66
18	33	56	2.5	79
19	29	46	1.9	88
20	33	49	2.1	60
21	55	51	2.4	49
22	29	52	2.3	77
23	44	58	2.9	52
24	43	50	2.3	60

The proposed model is

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

and the estimated regression equation is given by

$$\hat{y}_i = \underset{(6.88)}{155.11} - \underset{(-4.1)}{1.15 x_{i1}} + \underset{(-.5)}{.27 x_{i2}} - \underset{(-2.09)}{15.05 x_{i3}} \qquad (16)$$

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

For this equation, $R^2_{adj} = .634$, $F = 14.38$, $p < .001$. For Equation 16, X_2 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

Figure 2
Values of D_i, C_i, and h_i From Model 1 For Each Person

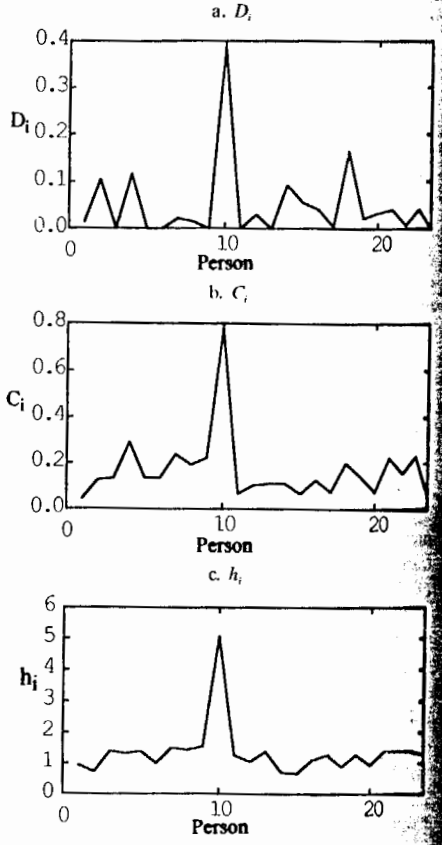


Table 2
Residuals, Distance, and Volume Regression Diagnostics for Models 1 and 2

Model and Person	Residuals			Distance		Volume	
	OLS	t_i	t_i^*	D_i	h_i	W_i	C_i
Model 1							
1	11.567	1.170	1.182	.0181	.0503	.272	.973
2	16.184	1.707	1.800	.1048	.1258	.683	.748
3	.209	.022	.022	.0001	.1339	.009	1.418
4	-9.172	-1.073	-1.077	.1172	.2895	-.688	1.363
5	-.493	-.051	-.051	.0001	.1375	-.020	1.423
6	-14.398	-1.458	-.051	.0001	.1375	-.020	1.423
7	-4.768	-.540	-.530	.0233	.2422	-.300	1.528
8	4.659	.513	.503	.0163	.1981	.250	1.452
9	.817	.091	.089	.0006	.2234	.048	1.578
10	3.178	.665	.655	.3866	.7778	1.226	5.053
11	2.771	.283	.276	.0014	.0678	.075	1.296
12	-9.845	-1.025	-1.026	.0301	.1029	-.348	1.103
13	-1.462	-.153	-.150	.0008	.1174	-.055	1.385
14	-16.629	-1.740	-1.840	.0950	.1115	-.652	.717
15	-16.785	-1.717	-1.813	.0565	.0712	-.502	.698
16	10.279	1.085	1.091	.0432	.1279	.418	1.104
17	-4.549	-.467	-.458	.0046	.0770	-.132	1.273
18	14.430	1.595	1.664	.1635	.2044	.844	.895
19	7.120	.758	.750	.0239	.1425	.306	1.274
20	-12.480	-1.280	-1.303	.0340	.0766	-.375	.944
21	6.769	.758	.749	.0414	.2240	.403	1.408
22	3.759	.404	.395	.0076	.1575	.171	1.410
23	6.586	.740	.731	.0408	.2297	.400	1.426
24	2.252	.229	.223	.0008	.0592	.056	1.291
Model 2							
1	12.132	1.214	1.231	.0224	.0573	.303	.953
2	13.858	1.550	1.614	.1946	.2448	.919	.957
3	.606	.063	.062	.0002	.1374	.025	1.438
4	-7.675	-.917	-.913	.1078	.3387	-.654	1.566
5	-1.777	-.190	-.185	.0019	.1738	-.085	1.491
6	-13.799	-1.383	-1.420	.0304	.0598	-.358	.864
7	-6.634	-.781	-.773	.0714	.3188	-.529	1.599
8	4.150	.452	.442	.0131	.2038	.224	1.493
9	1.550	.172	.168	.0023	.2352	.093	1.613
10	2.449	.247	.241	.0012	.0701	.066	1.318
11	-11.863	-1.283	-1.307	.0982	.1926	-.638	1.070
12	-.589	-.062	-.060	.0001	.1342	-.024	1.433
13	-14.310	-1.585	-1.656	.1873	.2298	-.904	.914
14	-16.954	-1.710	-1.810	.0567	.0719	-.504	.685
15	13.179	1.545	1.608	.2720	.3130	1.086	1.056
16	-3.449	-.354	-.346	.0036	.1037	-.118	1.349
17	14.888	1.627	1.707	.1749	.2091	.878	.861
18	7.220	.758	.749	.0239	.1428	.306	1.281
19	-12.219	-1.237	-1.255	.0324	.0781	-.366	.963
20	7.324	.812	.804	.0494	.2308	.440	1.402
21	3.660	.388	.379	.0070	.1577	.164	1.428
22	5.978	.666	.656	.0346	.2379	.366	1.482
23	2.273	.228	.222	.0008	.0592	.056	1.305

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

$$\hat{y}_i = \underset{(6.32)}{162.88} - \underset{(-4.01)}{1.21x_{i1}} - \underset{(-.81)}{.67x_{i2}} - \underset{(-.7)}{8.61x_{i3}} \quad , \quad (17)$$

which resulted in $R^2_{adj} = .621$, 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_j 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_js for Model 2 were 1.4, 2.8, and 2.9. The correlation between X_2 and X_3 was approximately .80, indicating strong colinearity. Thus, the removal of a single data point can reveal considerable multicolinearity. Extended discussions of multicolinearity from insertion and deletion of a small set of data points can be found in Chatterjee & Hadi (1988).

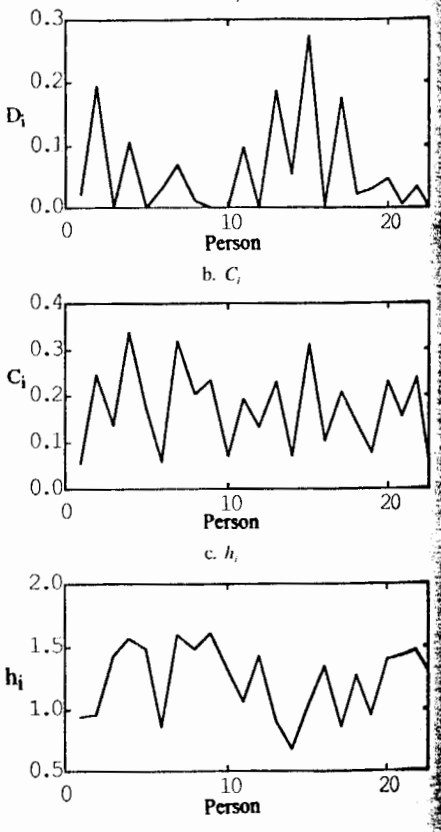
The diagnostics after deleting Person 10 also are given in Table 2, and plots of h_i, W_i, and C_i by person for this reduced dataset are given in Figure 3; these results do not reveal any influential data point.

If X_2 or X_3 is dropped, the regression equations are

$$\hat{y}_i = \underset{(6.69)}{166.59} - \underset{(-4.36)}{1.26x_{i1}} - \underset{(1.98)}{1.09x_{i2}} \quad , \quad (18)$$

with $R^2_{adj} = .631$

Figure 3
Values of D_i, C_i, and h_i From Model 2 For Each Person



and

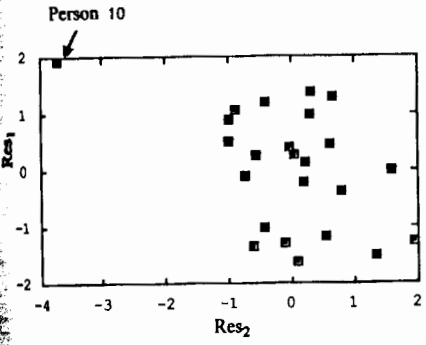
$$\hat{y}_i = \underset{(8.79)}{147.08} - \underset{(-4.20)}{1.243x_{i1}} - \underset{(-1.92)}{8.256x_{i3}} \quad , \quad (19)$$

with $R^2_{adj} = .627$; the diagnostic values showed no influential points.

Partial Residual Plot

The partial residual plot for X_3 was constructed by regressing Y on X_1 and X_2 and retrieving the residuals (called Res_1), and by regressing X_3 on the regressors X_1 and X_2 and retrieving the residuals (called Res_2). Figure 4 is the partial leverage plot for X_3 with Res_1 (vertical axis) against Res_2 (horizontal axis). Note the influential nature of Person 10. Without Person 10, the statistical significance of X_3 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 misspecification 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 explanatory 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 diagnostics 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 researcher can use the structural equations approach or study the sensitivity of the regression coefficients by any three of the methods mentioned earlier.

Discussion

Do regression diagnostics help uncover substantive issues? For the full dataset, the variability in Y was explained by X_1 and X_3, but X_2 was not statistically significant. However, when the data for Person 10 were dropped from the analysis, both X_2 and X_3 were jointly not significant. These results suggest that X_1 and either X_2 or X_3 should be used to explain the variability in Y. This parsimonious description of the data was facilitated through the use of regression diagnostics.

Such findings are not limited to small datasets (Belsley et al., 1980). Huber (1981) investigated 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 investigations have shown that the breakdown point of least squares estimates approaches 0 as the number of observations becomes large. Consequently, even a small percentage of influential points in large datasets can affect the results, emphasizing 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 exert influence on the results proportional to its weight. A weight of 0 implies deletion, and a weight of 1 leaves the observation in the estimation 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.

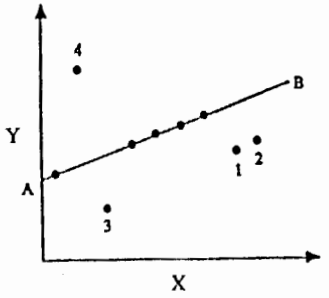
Although commonly available regression diagnostics are adequate in the context of linear regression models, there are some situations in

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

Table 3
Regression Diagnostics for Masked Data

Observation	x	y	D_i	h_i	t_i^*
1	7	13	.031	.27	-.38
2	8	14	.108	.41	.53
3	2	2	.491	.22	-2.51
4	2	14	.549	.22	2.84
5	1	5	.073	.13	-.52
6	6	15	.046	.18	.63
7	3	9	0.000	.14	-.03
8	5	13	.012	.13	.39
9	4	11	.002	.11	.17

$$\hat{y} = \underset{(2.18)}{5.28} + \underset{(2.53)}{1.28}x \quad . \tag{20}$$

h_i, D_i, and externally studentized residuals t_i^* also are shown in Table 3. Note that only Observations 3 and 4 are influential data points when measured by D_i and h_i. To demonstrate that Points 1 and 2 are jointly influential, Observation 1 was deleted, which resulted in the following regression equation:

$$\hat{y} = \underset{(1.91)}{5.05} + \underset{(2.31)}{1.37}x \quad , \tag{21}$$

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

$$\hat{y} = \underset{(1.7)}{4.71} + \underset{(2.26)}{1.48}x \quad . \tag{22}$$

However, if both Observations 1 and 2 are deleted,

$$\hat{y} = \underset{(1.16)}{3.72} + \underset{(2.15)}{1.87}x \quad . \tag{23}$$

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.

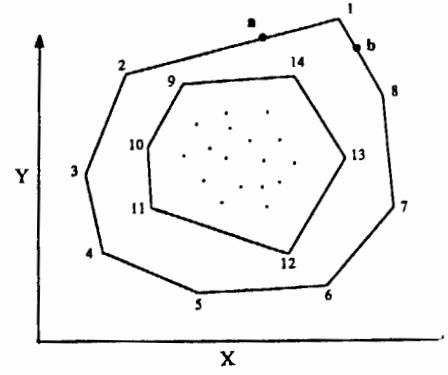
When masking is present, jointly influential points can be detected by computing the diagnostics after removing pairs, triplets, and so forth of points from the dataset. However, for any reasonable dataset, the number of computations become unmanageable. There are several strategies to deal with this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

Figure 6
The Outermost Convex Hull of a Set of Points



successfully by Chatterjee, Jamieson, & Wisema (1991) for detecting multiple influential points factor analysis. Peeling of convex hulls for ordering multivariate data can be found in Green (198 and other uses of the vertices of convex hulls extreme points are discussed in Barnett (1976).

Diagnostics for Multivariate Regression and Other Models

Multivariate Diagnostics

There appears to be very little work done multivariate diagnostics, other than the work Hossain & Naik (1989). Residual-based, volum based, and distance-based diagnostics for th multivariate regression model are discussed brie ly. There appears to be no commonly availabl software to implement these, but the diagnosti described below can be computed using the PR MATRIX command in SAS (SAS Institute, In 1982).

Consider the multivariate regression mode

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \quad , \tag{2}$$

where **Y** is a $n \times p$ response matrix of n obse vations of p response variables,
X is a $n \times (k + 1)$ matrix of regressor
B is a $(k + 1) \times p$ matrix of paramet 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 $\mathbf{e}_i'$ of $\mathbf{E}$ are assumed to be independent and normally distributed, with a $p \times 1$ mean vector of 0s and a $p \times p$ covariance matrix Σ. The least squares estimates of b_j are given by

$$b_j = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{Y}_j \qquad (25)$$
$$j = 1, 2, \ldots, p,$$

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 ith data point may be an influential point.

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

$$\hat{\Sigma} = \frac{\hat{\mathbf{E}}'\hat{\mathbf{E}}}{n - k - 1} \quad . \qquad (26)$$

Let

$$\hat{\Sigma}_i = \frac{\hat{\mathbf{E}}_{(i)}'\hat{\mathbf{E}}_{(i)}}{n - k - 1} \quad , \qquad (27)$$

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

$$r_i^2 = \frac{1}{(1 - h_i)}\mathbf{e}_i'\hat{\Sigma}^{-1}\mathbf{e}_i, \qquad (28)$$

and

$$T_i^2 = \frac{1}{(1 - h_i)}\mathbf{e}_i'\hat{\Sigma}_{(i)}^{-1}\mathbf{e}_i \quad , \qquad (29)$$

where r_i^2 and T_i^2 are the multivariate analogues of t_i and t_i^* 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

$$D_i^* = \frac{r_i^2}{k + 1}\frac{h_i}{1 - h_i}, \quad i = 1, 2, \ldots, n. \qquad (30)$$

Influence of the ith observation on the estimate of $\hat{\mathbf{B}}$ is indicated by large values of D_i^*.

W_i can be modified accordingly as

$$W_i^* = \frac{h_i}{1 - h_i} T_i^2 \quad , \qquad (31)$$
$$i = 1, 2, \ldots, n,$$

and a cutoff value is given by

$$\frac{k + 1}{n}\frac{p(n - k - 2)}{n - k - p - 1} F_{\alpha,p,n-k-p-1} \quad , \qquad (32)$$

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^* is the corresponding multivariate analogue of the scalar variances and is given by

$$C_i^* = \left(\frac{1}{1 - h_i}\right)^p \left(\frac{||\hat{\Sigma}||}{||\hat{\Sigma}_i||}\right)^{k+1} \quad . \qquad (33)$$

As before, very high or very low values of C_i^* 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 Jørgenson (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_i 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_i, Cook's distance; (4) W_i, the Welsch-Kuh distance; and (5) C_i, 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 multicolinearity, 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 are

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

Atkinson, A. C. (1985). *Plots, transformations, and regression.* Oxford, England: Oxford University Press

Barnett, V. (1976). Ordering multivariate data (with discussion). *Journal of the Royal Statistical Society, 139,* (Series A), 318–354.

Belsley, D. A. (1984). Demeaning condition diagnostics through centering (with discussion). *The American Statistician, 38,* 73–93.

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

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K. A., & Jackman, R. W. (1985). Regression diagnostics: An expository treatment of outliers and influential cases. *Sociological Methods and Research, 13,* 510–542.

Box, G. E. P. (1983). Apology for eclumenicism in statistics. In G. E. P. Box & C. F. Wu (Eds.), *Scientific inference and data analysis* (pp. 121–145). New York: Wiley.

Casella, G. (1983). Leverage and regression through the origin. *The American Statistician, 37,* 147–151.

Chatterjee, S., & Chatterjee, S. (1990). Convex hull of a set of points. *Computational Statistics and Data Analysis, 10,* 87–92.

Chatterjee, S., & Hadi, A. S. (1986). Influential observations, high leverage points and outliers in linear regression (with discussion). *Statistical Science, 1,* 379–415.

Chatterjee, S., & Hadi, A. S.(1988). *Sensitivity in linear regression.* New York: Wiley.

Chatterjee, S., Jamieson, L., & Wiseman, F. (1991). Diagnostics in factor analysis. *Marketing Science, 10,* 145–160.

Chatterjee, S., & Price, B. P. (1973). *Regression analysis by example.* New York: Wiley.

Chatterjee, S., & Wiseman, F. (1984). Use of regression diagnostics in political science research. *American Journal of Political Science, 27,* 601–613.

Chesher, A. (1991). The effect of measurement error. *Biometrika, 78,* 451–462.

Cochran, W. G. (1972). Some effects of errors of measurement on linear regression. In J. Neyman (Ed.), *Proceedings of the 6th Berkeley Symposium, 1,* 527–539.

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, 46,* (Series B), 133–169.

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

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

Cook, R. D., & Weisberg, S. (1982b). Criticism and influence analysis in regression. In S. Leinhardt (Ed.), *Sociological Methodology* (pp. 313–361).

Daniel, C., & Wood, F. S. (1980). *Fitting equations to data: Computer analysis of multifactor data* (2nd ed.). New York: Wiley.

Darlington, R. B. (1990). *Regression and linear models.* New York: McGraw Hill.

Delaney, N. J., & Chatterjee, S. (1986). Use of the bootstrap and cross-validation in ridge regression. *Journal of Business and Economics Statistics, 4,* 255–261.

Dixon, W. J. (Ed.). (1984). *BMDP statistical software.* Berkeley CA: University of California Press.

Fuller, W. A. (1987). *Measurement error models.* New York: Wiley.

Gray, J. B., & Ling, R. F. (1984). K-Clustering as a detection for influential subsets in regression (with discussion). *Technometrics, 26,* 305–330.

Green, P. J. (1984). Peeling of bivariate data. In V. Barnett (Ed.), *Interpreting multivariate data* (pp. 235–271). New York: Wiley.

Heller, G. (1987). *Fixed sample results for linear models with measurement error.* Unpublished doctoral dissertation, New York University.

Hossain, A., & Naik, D. N. (1989). Detection of influential observations in multivariate regression. *Journal of Applied Statistics, 16,* 21–32.

Huber, P. (1981). *Robust statistics.* New York: Wiley.

Huber, P. J. (1983). Minimax aspects of bounded-influence regression (with comments). *Journal of the American Statistical Association, 78,* 66–80.

Jørgenson, B. (1983). Maximum likelihood estimation and large sample inference by generalized linear and nonlinear regression models. *Biometrika, 70,* 19–28.

Judge, G. G., Griffith, W. L., Hill, R. C., & Lee, T. C. (1980). *The theory and practice of econometrics.* New York: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Malinvaud, E. (1970). *Statistical methods in econometrics.* Amsterdam: North Holland.

Mansfield, E. R., & Helms, B. P. (1982). Detecting multicollinearity. *The American Statistician, 36,* 158–160.

McCullagh, P., & Nelder, J. A. (1984). *Generalized linear models.* New York: Chapman and Hall.

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

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw Hill.

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

Picard, R. R., & Berk, K. N. (1990). Data splitting. *The American Statistician, 44,* 140–147.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association, 79,* 575–583.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics, 9,* 705–724.

Polasek, W. (1984). Regression diagnostics for general linear regression models. *Journal of the American Statistical Association, 79,* 79–86.

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

Rousseeuw, P. J., & LeRoy, A. N. (1987). *Robust regression and outlier detection.* New York: Wiley.

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

Ryan, B. F., Joiner, B. L., & Ryan, T. A. (1991). *MINITAB handbook* (2nd ed.). Boston MA: Duxbury Press.

SAS Institute, Inc. (1982). *SAS user's guide: Basics. Cary NC: Author.*

SPSS, Inc. (1988). SPSS-X user's guide (3rd ed.). Chicago IL: Author.

Stefanski, L. A. (1985). The effect of measurement error on parameter estimation. *Biometrika, 72,* 583–592.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin, 95,* 334–344.

Stewart, G. W. (1987). Collinearity and least squares regression. *Statistical Science, 2,* 68–100.

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

Wellman, J. M., & Gunst, R. F. (1991). Influence diagnostics for linear measurement error models. *Biometrika, 78,* 373–380.

Welsch, R. E., & Kuh, E. (1977). *Linear regression diagnostics* (Tech. Rep. No. 923-77). Cambridge: Massachusetts Institute of Technology, Sloan School of Management.

Wilkinson, L. (1992). *SYSTAT: Statistics* (Version 5.2). Evanston IL: SYSTAT, Inc.

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.