26
5-3-82
J. C. Daly
NTIS

I-2857

(1)

Lh. 460

PNL-4208
UC-32

MASTER

DO NOT MICROFILM
COVER

Diagnostics for Multiple-Regression Problems

J. C. Daly

March 1982

Prepared for the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Operated for the U.S. Department of Energy
by Battelle Memorial Institute

PNL-4208

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC06-76RLO 1830

Printed in the United States of America
Available from
National Technical Information Service
United States Department of Commerce
5285 Port Royal Road
Springfield, Virginia 22151

NTIS Price Codes
Microfiche A01

Printed Copy

Pages	Price Codes
001-025	A02
026-050	A03
051-075	A04
076-100	A05
101-125	A06
126-150	A07
151-175	A08
176-200	A09
201-225	A010
226-250	A011
251-275	A012
276-300	A013

PNL--4208

DE82 013629

DIAGNOSTICS FOR
MULTIPLE-REGRESSION PROBLEMS

J. C. Daly

DISCLAIMER

This book was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

March 1982

Prepared for
the U.S. Department of Energy
under Contract DE-AC06-76RLO 1830

Pacific Northwest Laboratory
Richland, Washington 99352

cb
DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

ACKNOWLEDGMENTS

This work would not have been possible without the joint sponsorship of the Analysis of Large Data Sets project sponsored by the Applied Mathematics Section of Basic Energy Sciences at DOE and a grant from the NSF Science Faculty Professional Development Program. The work was completed during a 10-week visiting faculty assignment at Pacific Northwest Laboratory.

Anthony R. Olsen, Wes Nicholson, Dan Carr and Dave Hall spent many hours in discussions on data analysis philosophy and specific points on diagnostic procedures. Unknowingly, Daryl Pregibon, Sanford Weisberg and Dennis Cook contributed their ideas to the project during discussions with them. Without the patience of Sharon Popp, who typed the manuscript, this report would not have been completed.

THIS PAGE
WAS INTENTIONALLY
LEFT BLANK

CONTENTS

1.	INTRODUCTION	1
2.	FORMULATION AND DEFINITION OF SINGLE CASE REGRESSION DIAGNOSTICS	3
2.1	THE HAT MATRIX	3
2.2	CHANGE IN SLOPE COEFFICIENTS WHEN DELETING SINGLE OBSERVATION	5
2.3	CALCULATIONS ON THE STANDARD DEVIATION BASED ON REDUCED X MATRIX	6
2.4	"STANDARDIZED" RESIDUALS	7
2.5	MEASURES OF EFFECTS OF INDIVIDUAL OBSERVATIONS ON ESTIMATED $\underline{Y}$, $\hat{\underline{Y}}$	9
2.6	DIAGNOSTIC MEASURES BASED ON DETERMINANTS	12
2.7	EFFECTS ON INDIVIDUAL PARAMETERS	14
2.8	PARTIAL REGRESSION (RESIDUAL) PLOTS	16
2.9	INFLUENCE OF SUBSET OF β	17
2.10	DIAGNOSTICS FOR WEIGHTED LEAST SQUARES	17
3.	IMPLEMENTATION	19
3.1	LEVEL I IMPLEMENTATION	19
3.2	LEVEL II IMPLEMENTATION	32
3.3	LEVEL III IMPLEMENTATION	33
4.	ANALYSIS OF EXAMPLE DATA	35
4.1	THE FITTED MODEL	36
4.2	ANALYSIS OF HAT MATRIX DIAGONALS	36
4.3	ANALYSIS OF "STANDARDIZED" RESIDUALS	39
4.4	ANALYSIS OF EFFECT ON $\hat{Y}_i$	42
4.5	DIAGNOSTICS BASED ON DETERMINANTS	46

4.6	EFFECTS ON INDIVIDUAL PARAMETERS	50
4.7	PARTIAL REGRESSION PLOTS	59
4.8	INFLUENCE OF SUBSET OF β	63
4.9	SUMMARY	64
	BIBLIOGRAPHY	66
	APPENDIX A - 1969 POLLUTION-MORTALITY DATA	A.1
	APPENDIX B - LISTING OF SMSA CODE NUMBERS AND LOCATIONS FOR 1969 POLLUTION-MORTALITY DATA	B.1
	APPENDIX C - LISTING OF DIAGNOSTICS	C.1
	APPENDIX D - LISTING OF DFBETAS	D.1
	APPENDIX E - LISTING OF COOK'S $D_i(\hat{\beta}_j)$	E.1
	APPENDIX F - LISTING OF $D_i(\beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11})$	F.1
	APPENDIX G - QR DECOMPOSITION	G.1

FIGURES

1	Regression Output for Example	37
2	Stem-and-leaf Display for h_i	38
3A	Stem-and-leaf Display for Standardized Residuals, t_i	40
3B	Stem-and-leaf Display for Studentized Residuals, r_i^*	40
4	Half-normal Plot for $ r_i^* $	41
5A	Stem-and-leaf Display for DFFITS	43
5B	Stem-and-leaf Display for D_i	43
5C	Stem-and-leaf Display for T_i	44
6	Half-normal Plot of T_i	45
7	Stem-and-leaf Display of Andrews-Pregibon $R_i^1(X^*)$	48
8	Stem-and-leaf Display for $COVRATIO_i$	49
9	Stem-and-leaf Display for Cook's Volume Ratio	51
10A	Plot of $D_i(\hat{\beta}_2)$ vs. Observational Number	54
10B	Plot of $DFBETAS_{i2}$ vs. Observational Number	54
10C	Plot of Percent Change in $\hat{\beta}_2$ vs. Observational Number	58
11A	Plot of $D_i(\hat{\beta}_7)$ vs. Observational Number	56
11B	Plot of $DFBETAS_{i7}$ vs. Observational Number	56
11C	Plot of Percent Change in $\hat{\beta}_7$ vs. Observational Number	58
12A	Plot of $D_i(\hat{\beta}_8)$ vs. Observational Number	57
12B	Plot of $DFBETAS_{i8}$ vs. Observational Number	57
12C	Plot of Percent Change in $\hat{\beta}_8$ vs. Observational Number	59
13	Partial Regression Plot Associated with $\hat{\beta}_2$	60
14	Partial Regression Plot Associated with $\hat{\beta}_7$	61
15	Partial Regression Plot Associated with $\hat{\beta}$	62

16	Stem-and-leaf Display for $D_i(\psi)$	64
17	Plot of Hat Matrix Diagonals for Reduced Model $(\beta_1 - \beta_0)$ vs. Hat Matrix Diagonals for Full Models	65

TABLES

1	Leverage Points (Cutoff = 0.214)	38
2	Extreme Values for "Standardized" Residuals (Cutoff = ± 2)	42
3	Points of Influence for DFFITS, D_i and T_i (Cutoff Value: $ DFFITS > 0.655$)	46
4	Extreme Values for Andrews-Pregibon $R_i^1(X^*)$	48
5	Extreme Values for COVRATIO	49
6	Extreme Values for Cook's Volume Ratio	51
7	Summary of Occurrences of DFBETAS Exceeding Cutoff Values	53
8	Listing of SMSAs and Number of Parameters Where DFBETAS Exceeds $2/\sqrt{n}$	53

DIAGNOSTICS FOR MULTIPLE REGRESSION PROBLEMS

1. INTRODUCTION

In the last 10 to 15 years there has been much work done in trying to improve linear regression results. Individuals have analyzed the susceptibility of least-squares results to values far removed from the center of the independent variable observations. They have studied the problem of heavy-tailed residuals, and they have studied the problem of collinearity. From these studies have come ridge regression techniques, robust regression techniques, regression on principal components, etc. However, many practitioners view these methods with suspicion (and ignorance), and prefer to continue using the usual least-squares procedures to fit their models, even though their results might not be answering the question they think.

In reaction to this, statisticians are spending more time analyzing how the individual observations affect the least squares results. In the last few years approximately 10 papers and one text have appeared that address the problem of how to study the influence of the individual observations.

This report is a study of the recent work done in linear regression diagnostics. It will be concerned with analyzing the effect of one case at a time, since the methods to analyze this situation are relatively straightforward and are not prohibitive computationally.

The regression diagnostic techniques discussed have recently been added to a few widely available statistical computing packages. Consequently, diagnostic techniques will be used as a standard part of many regression analyses. However, a detailed study of a regression fitting procedure introduces a number of statistical computing problems. Foremost is the design of a user friendly interactive module to allow the statistician to thoroughly complete a data analysis. A batch oriented computer program no longer suffices. A second consideration is the completion of the diagnostic analysis on large data sets. Here the primary emphasis of large is the complexity of the problem. Generally, a reasonably large number of regressor variables are

available for inclusion in the regression. The selection of the variables and the completion of diagnostics for every regression fitted results in a large number of computations and taxes the data analyst's ability to remember past results. Considerable work on both of these problems is needed before a data analyst can routinely complete a detailed regression analysis on a complex (large) data set. This report is the first step toward studying these problems.

The report will consist of three main chapters. Chapter 2 is a discussion of the diagnostics that are available, with some attempt made to motivate them. Chapter 3 deals with implementation. In this chapter we look at the possible ways to implement part or all of the various regression diagnostics as part of an existing computer package. Also, we indicate the information from a regression program needed to implement a complete regression diagnostic package. Chapter 4 is a complete worked example, illustrating the diagnostic measures a step at a time. Individuals familiar with regression diagnostics might wish to follow the worked example as they read through the discussion of the various diagnostics.

One additional comment is necessary before the discussion of the diagnostics begin. The principal references used are a book by Belsley, Kuh and Welsch (1980) and a series of papers by Cook (1977, 1979) and Cook and Weisberg (1980). The notation is different in the two sources and the notation used in this report is a combination of the notation used in those and other sources. The notation, however, is consistent throughout, and the names or symbols given any diagnostics in the original papers have been maintained where possible.

2. FORMULATION AND DEFINITION OF SINGLE CASE REGRESSION DIAGNOSTICS

In the following discussion it is assumed that we are fitting the linear model, $\underline{Y} = \underline{X}\underline{\beta} + \underline{\epsilon}$, where $\underline{Y}$ is an $(n \times 1)$ column vector of observations on the dependent variable, $\underline{\beta}$ is a $(p \times 1)$ column vector of unknown parameters, $\underline{X}$ is an $(n \times p)$ matrix of observations on the independent variables, and $\underline{\epsilon}$ is an $(n \times 1)$ vector of random variables, where $\underline{\epsilon} \sim N_n(\underline{0}, \sigma^2 \underline{I})$.

This model is fit by the least-square process, resulting in a $(p \times 1)$ vector of parameter estimates $\hat{\underline{\beta}}$, where $\hat{\underline{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y}$. Let $\underline{r} = (\underline{Y} - \underline{X}\hat{\underline{\beta}}) = (\underline{Y} - \hat{\underline{Y}})$ be the vector of residuals, $RSS = \sum (Y_i - \hat{Y}_i)^2 = \sum r_i^2$, and let $\underline{x}_i'$ be the i th row of the $\underline{X}$ matrix (the values of the independent variables for the i th observation). Unless otherwise stated, it is assumed that the model contains a y -intercept term, making the first element of $\underline{x}_i'$ the value 1. The diagnostics considered in this chapter have been developed to determine which observations, if any, are having sizable effects on our estimated regression line. By an observation having "sizable effects," we refer to the situation where, depending on whether a particular observation is included or deleted, the regression results change significantly. The order in which the diagnostics are considered roughly corresponds to an arrangement of increasing independence on previously discussed diagnostics, and complexity of calculations.

For each topic, when appropriate, we give the definition and rationale for the diagnostic, the reduced form of the diagnostic (the form in which it is most readily computed), and discuss diagnostic criteria and displays.

2.1 THE HAT MATRIX

Hoaglin and Welsch (1978) defined $H = \underline{X}(\underline{X}'\underline{X})^{-1}\underline{X}'$ to be the hat matrix in least squares regression. The name is due to the fact that $\hat{\underline{Y}} = H\underline{Y}$. This matrix and its elements are very important in linear regression. In the first place, H is both symmetric and idempotent, and is a projection operator onto the range spanned by the columns of $\underline{X}$. In particular, the diagonal elements of this matrix play a very important part. If we let h_{ij} denote the value in the i th row, j th column of H , $h_{ii} = h_i$ will denote the i th diagonal element

of H. Belsley, Kuh and Welsch (1980, pp. 16, 66) mention several interesting facts concerning h_i . To understand these results, consider the relationship $\hat{\underline{Y}} = H\underline{Y}$. This implies that

$$\hat{y}_i = \sum_{j=1}^n h_{ij} y_j$$

From this we see that h_i is the "weight" placed on the i th observation y_i in determining the i th estimated value, $\hat{y}_i$. However, since H is symmetric and idempotent, it is fairly straightforward to show that $1/n \leq h_i \leq 1$, and

$$h_i = \sum_{j=1}^n h_{ij}^2$$

Using these two facts it is obvious that the higher the value of h_i (closer to 1), the greater the "leverage" that the values of the independent variables for the i th data point have in determining $\hat{y}_i$. In fact, if $h_i = 1$, all other values in the i th row of the H matrix = 0, $\hat{y}_i = y_i$, and $r_i = 0$. The value h_i is also related to the distance from $\underline{x}_i$ to $\underline{\bar{x}}$, the vector of independent variable means, i.e., the larger h_i , the farther the distance.

Other results involving h_i concern $\text{Var}(\hat{y}_i)$ and $\text{Var}(r_i)$. We know that

$$\text{Var}(\hat{y}_i) = \underline{x}_i' (X'X)^{-1} \underline{x}_i \sigma^2$$

and

$$\text{Var}(r_i) = (1 - \underline{x}_i' (X'X)^{-1} \underline{x}_i) \sigma^2$$

But $\underline{x}_i' (X'X)^{-1} \underline{x}_i = h_i$, so

$$\text{Var}(\hat{y}_i) = h_i \sigma^2$$

and

$$\text{Var}(r_i) = (1 - h_i) \sigma^2$$

The reduced form for the values of the H matrix are usually given in the form mentioned with respect to $\text{Var}(\hat{y})$; namely

$$h_{ij} = \underline{x}_i' (X'X)^{-1} \underline{x}_j$$

and

$$h_i = \underline{x}_i' (X'X)^{-1} \underline{x}_i .$$

Since H is idempotent, $\text{tr}(H) = \sum h_i = r(H) = p$. Therefore, an "average" value of h_i is p/n . Belsley, Kuh and Welsch (1980) chose $2p/n$ as an arbitrary breaking point. If $h_i > 2p/n$, the i th data point is called a "leverage" point, indicating that data point might have a large effect on our regression line. The value $2p/n$ should be used judiciously. If p is large relative to n ($p/n > 0.4$), leverage points will go undetected, while if p is very small relative to n , a large number of points may exceed the cutoff value. This second situation might require more time on the analyst's part, but shouldn't result in any true "points of influence" being ignored.

The values of h_i could be displayed in a number of ways. One possibility is a stem-and-leaf display. Another way is plotting h_i against case number. Plotting h_i against each of the independent variables would be time-consuming but might show the analyst why a certain row of X has a high value.

In all of the above discussion regarding the hat matrix and its diagonal elements, nothing has been said regarding the y values, or the effect that a row of the X matrix has on our estimated regression equation. The hat matrix as a diagnostic tool merely helps us identify observations that might have a large effect on the regression equation, and gives us a building block on which to base other diagnostic tools.

2.2 CHANGE IN SLOPE COEFFICIENTS WHEN DELETING SINGLE OBSERVATION

In determining the effect an observation has on the estimated regression equation, one logical starting point is to consider the change in the estimated parameters if we delete the values of the observation in question $(\underline{x}_i', y_i)$. Let $\underline{y}_{(-i)}$ be the vector of observations on the dependent variable

with the i th element deleted. Let $X_{(i)}$ be the X matrix with the i th row deleted. We want to consider $\hat{\beta} - \hat{\beta}_{(i)}$, where $\hat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}Y_{(i)}$. By writing $\hat{\beta}_{(i)} = (X'X - x_i x_i')^{-1}(X'Y - x_i y_i)$, and expanding the inverse, we can determine the following reduced form of the relationship:

$$\hat{\beta} - \hat{\beta}_{(i)} = \frac{(X'X)^{-1}x_i r_i}{(1 - h_i)} \quad (1)$$

(Belsley, Kuh and Welsch, 1980, pp. 13, 64; Cook 1977, p. 16). This relationship doesn't serve as much of a diagnostic by itself, but it is a basic part of many others. The formula does indicate when there will be large differences between $\hat{\beta}$ and $\hat{\beta}_{(i)}$. If both r_i and h_i are large, the difference will be large. If both are small, the difference will be small. When one term is small and the other large, other factors will determine the magnitude of the values of $\hat{\beta} - \hat{\beta}_{(i)}$.

2.3 CALCULATION OF THE STANDARD DEVIATION BASED ON REDUCED X MATRIX

From the model with all observations, the least squares estimate of σ^2 is

$$s^2 = \sum (y_j - \hat{y}_j)^2 / (n - p) = \text{RSS} / (n - p)$$

If one observation is deleted, define

$$\begin{aligned} s^2(i) &= \sum_{j \neq i} (y_j - \hat{y}_j(i))^2 / (n - p - 1) \\ &= \text{RSS}(i) / (n - p - 1) \end{aligned}$$

Belsley, Kuh, and Welsch (1980, p. 64), hereafter referred to as BKW, show that

$$\text{RSS}(i) = \text{RSS} - r_i^2 / (1 - h_i)$$

From this result it follows that $s^2(i)$ can be expressed in the following reduced form:

$$s^2(i) = \left(\frac{1}{n-p-1}\right) \left[\text{RSS} - \frac{r_i^2}{(1-h_i)} \right] \quad (2)$$

In regression diagnostics both s^2 and $s^2(i)$ are used, with various authors having their particular favorites. The implications of the choice of error variance is discussed with the next diagnostics.

A plot of $s(i)$ versus i will indicate which points have relatively large residuals, and which points have small residuals. At an observation where the residual is relatively high, $s(i)$ will be smaller than s . At observations where the residual is relatively low, $s(i)$ will be larger than s . The exact meaning of "relatively" is given in the next section.

2.4 "STANDARDIZED" RESIDUALS

When analyzing residuals for possible outliers, and when plotting residuals to check on the existence of various patterns (heteroscedasticity, nonlinearity, etc.), it is common practice to divide each residual by an estimate of its standard deviation. In regression diagnostics, there are two different estimates available. Most computer packages calculate $t = r_i / \text{s.e.}(r_i)$, and call this term the standardized residual. It is straightforward to derive the following reduced form:

$$t_i = r_i / (s\sqrt{1-h_i})$$

Cook (1977, 1979) and Cook and Weisberg (1980) use this in their diagnostic methods. Another possibility in regression diagnostics is to consider the fitted model based on $(n-1)$ observations (omitting observation i). Let $\hat{y}_{(i)} = \mathbf{x}_i' \hat{\beta}_{(i)}$ be the predicted value of y_i at the new independent variable data point. Then a natural standardization for determining agreement between y_i and $\hat{y}_{(i)}$ is

$$r_i^* = \frac{y_i - \hat{\beta}(i)}{s(i) \sqrt{1 + \underline{x}_i' (X_{(i)}' X_{(i)})^{-1} \underline{x}_i}} \quad **$$

But, by various manipulations it can be shown that

$$y_i - \hat{\beta}(i) = r_i / (1 - h_i)$$

and

$$\underline{x}_i' (X_{(i)}' X_{(i)})^{-1} \underline{x}_i = h_i / (1 - h_i)$$

From these results the reduced form for r_i^* is

$$r_i^* = r_i / (s(i) \sqrt{1 - h_i})$$

This expression will be called the studentized residual. Following the argument of Atkinson (1981, p. 14) a case can be made for the use of r_i^* over t_i . If our data contains no outliers, the two sets of "standardized" residuals should be about the same. However, if there are outliers, the values of $|r_i|$ will be large, the values of $s(i)$ will be smaller than s , and $|r_i^*|$ will be greater than $|t_i|$. The statistic r_i^* thus has the property of "inflating" residuals at points that have large residuals. In fact, it is straightforward to show that $|r_i^*| > |t_i|$ if $|r_i| > s\sqrt{1 - h_i}$. This means any residual that is more than one standard error away from zero will be amplified by using r_i^* in place of t_i . The statistic r_i^* is used by both BKW and Atkinson. Atkinson calls the r_i^* the jackknife residuals, and points out that if the i th observation is an outlier (i.e., is not correctly modeled by the model under study), then $s^2(i)$ will still be an unbiased estimator of σ^2 , while s^2 won't be unbiased.

The two statistics t_i and r_i^* are both t -like statistics, and various percentage points from the t -Distribution can be used as approximate measures of extremeness for the residuals. In fact, r_i^* , under fairly general conditions does possess a t -Distribution with $(n-p-1)$ d.f. However, care must be taken

** The notation in BKW is e_i^* .

in doing any tests since the r_i are not independent. Many papers in the recent literature have dealt with methods of solution for the problem of outliers and we will not discuss them here (Tietjen, Moore and Beckman, 1973; Prescott, 1975; Lund, 1975; Ellenberg, 1976, and Cook and Prescott, 1981).

Many displays can be made using these statistics. Histograms and stem-and-leaf plots can be used to isolate the extreme values. Plots of t_i or r_i^* versus h_i can be done to show how the regression model fits at the leverage points. Atkinson (1981, p. 15) recommends plotting r_i^* as half-normal plots, where the ordered values are plotted on normal probability paper on the horizontal line given by $50 + 50(i - 1/2)/n$. (See Draper and Smith (1981, pp. 179-182) for further discussion on half-normal plots.)

2.5 MEASURES OF EFFECTS OF INDIVIDUAL OBSERVATIONS ON ESTIMATED $\underline{y}$, $\hat{\underline{y}}$

Nearly everything considered so far has done little to show the actual effects that certain observations have on our estimated regression equation. In this section, three closely related measures will be considered.

1. First consider the approach taken by Belsley, Welsch and Kuh (1980, p. 15). At the i th point, the difference between the estimated value with all observations used ($\underline{x}_i' \hat{\beta}$) and the estimated value where the i th row of X is omitted ($\underline{x}_i' \hat{\beta}_{(i)}$) is

$$\hat{y}_i - \hat{y}_{(i)} = \underline{x}_i' (\hat{\beta} - \hat{\beta}_{(i)}) = \frac{\underline{x}_i' (X'X)^{-1} \underline{x}_i r_i}{(1-h_i)} = \frac{h_i r_i}{(1-h_i)}.$$

In order to scale this difference, they suggest dividing by the term $s(i)\sqrt{h_i}$, a standard error for y_i (using $s(i)$ instead of s to estimate σ). This yields the following reduced form:

$$DFFITS = \left[\frac{h_i}{1-h_i} \right]^{1/2} \frac{r_i}{s(i)\sqrt{1-h_i}} = r_i^* \left[\frac{h_i}{1-h_i} \right]^{1/2}. \quad (3)$$

This is one of the principal diagnostic tools used by BKW.

2. The second measure of the effect of individual observations on our regression line is due to Cook (1977). Cook defines a distance measure that is a function of what he calls the empirical influence function for $\hat{\beta}$ ($\hat{\beta}_{(i)} - \hat{\beta}$), a positive definite matrix M and a nonzero scaling factor c , i.e.,

$$D_i(M, c) = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' M (\hat{\beta}_{(i)} - \hat{\beta})}{c}$$

Cook mentions many different choices for M and c , but prefers $M = X'X$ and $c = ps^2$. For this choice

$$\begin{aligned} D_i(X'X, ps^2) &= D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X'X (\hat{\beta}_{(i)} - \hat{\beta})}{ps^2} \\ &= \frac{(\hat{X}\hat{\beta}_{(i)} - \hat{X}\hat{\beta})' (\hat{X}\hat{\beta}_{(i)} - \hat{X}\hat{\beta})}{ps^2} \\ &= \frac{(\hat{Y}_{(i)} - \hat{Y})' (\hat{Y}_{(i)} - \hat{Y})}{ps^2} \end{aligned}$$

This last term is the scaled sum of squares of the changes of the estimated y -values caused by deleting the i th observation. A large value of D_i indicates a big change in the estimated Y values, which is one of the characteristics of a point of influence.

By using our basic identity for $\hat{\beta} - \hat{\beta}_{(i)}$ (Eq. 1), D_i can be expressed in the following reduced form:

$$D_i = \left(\frac{1}{p}\right) t_i^2 \left(\frac{h_i}{1-h_i}\right) \quad (4)$$

The value t_i^2 is measuring the effect of the residual, and the ratio $(h_i/(1-h_i))$ can be viewed as the ratio of the variance of $\hat{y}_i$ to the variance of r_i . Thus, D_i can be large when we have

a potential outlier (large t_i^2) and/or we have a high leverage point (h_i large). Cook and Weisberg (1980, p. 497) give the reduced form of $D_i(M, c)$ for various choices of M and c . In fact, using $M = X'X$ and $c = ps^2(i)$ gives $D_i(M, c) = (\text{DFFITS})^2/p$. This relationship shows that DFFITS is a more general diagnostic than appeared to be the case when it was first defined. Instead of just being a measure of the agreement of the estimated values at the i th data point, it also measures the sum of squares of the changes in estimated values at all data points.

3. The third diagnostic, due to Atkinson, is merely a combination of the ideas expressed by BKW and Cook. Atkinson defines (1981, p. 15) the reduced form for his statistic as the following:

$$T_i = \left[\frac{n-p}{p} \left(\frac{h_i}{1-h_i} \right) \right]^{1/2} |r_i^*| \quad (5)$$

This is $|\text{DFFITS}|$ multiplied by $((n-p)/p)$. Atkinson does this to standardize the statistic for the effect of n against a D-optimum design where h_i is p/n for all i . He recommends the absolute value in place of the square used by Cook for ease of plotting.

The three statistics given in this section should all be used in a relative sense. There is no clean distribution theory to use in judging "unusual" values, but the various authors have come up with certain benchmarks. BKW (1980, p. 28) suggest $2\sqrt{p/n}$ as a cutoff value for DFFITS. They also suggest finding $\tilde{s}$, the interquartile range for the n values $\hat{y}_i - \hat{y}_{i(i)}$, $i = 1, \dots, n$, and flagging any observation where $|\hat{y}_i - \hat{y}_{i(i)}| > (7/2)\tilde{s}$ (p. 29). Cook and Weisberg suggest relative comparisons within the data set, and mention the fact that $F_i = t_i^2(n-p-1)/(n-p-t_i^2)$ follows an F-distribution with 1 and $(n-p-1)$ degrees of freedom (1980, p. 496). However, $F_i = (r_i^*)^2$, where r_i^* was defined in section 2.4, so the distributional result for F_i , given the necessary model assumptions, is expected. Atkinson (1981) doesn't

suggest any testing. For all three measures, a large value indicates more than usual influence by that point.

The usual means of displaying these statistics is in tabular form. However, a stem-and-leaf or a histogram could be used to identify the points of strong influence. Plots of any of these statistics against the h_i values can be used to show how much the location of the i th row of the X matrix in the independent variable hull (IVH) (Cook, 1979) contributes to the value of the statistic. Atkinson suggests half normal plots of the T_i , with "large values" causing the expected line to be bent upwards from a slope of 45° .

2.6 DIAGNOSTIC MEASURES BASED ON DETERMINANTS

As in the previous section, there are three diagnostic measures that fall under this heading. One is due to BKW, one is due to Andrews and Pregibon, and one is due to Cook and Weisberg.

1. Let us first consider the diagnostic due to BKW (1980, p. 22). They define

$$\text{COVRATIO} = \frac{\det\{s^2(i)[X'_{(i)}X_{(i)}]^{-1}\}}{\det\{s^2[X'X]^{-1}\}}$$

This statistic is merely the ratio of the determinant of the covariance matrix of the parameter estimates when the i th observation is not used to the determinant of the covariance matrix when all observations are used. The authors develop bounds for this ratio for the case where we have a potential outlier, and for the case where we have a leverage point. If $\text{COVRATIO} < 1 - 3/n$, the point is a potential outlier, and if $\text{COVRATIO} > 1 + 3/n$, there is a possibility of a point with extreme leverage. The reduced form for COVRATIO is:

$$\text{COVRATIO} = \left[\left(\frac{n-p-1}{n-p} + \frac{(r_i^*)^2}{n-p} \right)^p (1-h_i) \right]^{-1}$$

2. Andrews and Pregibon (1978, p. 88) look at the ratio of two determinants where they consider the X matrix to be augmented by the $\underline{Y}$ vector and designated as X^* . Then

$$R_i^1(X^*) = \frac{|X_{(i)}^{*'} X_{(i)}^*|}{|X^{*'} X^*|}$$

They argue that this ratio responds to both extreme points in the design space (IVH), and observations that appear to be outliers. Unlike most diagnostics previously considered, $R_i^1(X^*)$ attains small values (close to 0) when there are deviant and/or influential observations. The reduced form is the following:

$$R_i^1(X^*) = \left(1 - h_i\right) \left(\frac{n - p - t_i^2}{n - p}\right)$$

3. Cook and Weisberg (1980, p. 501) consider the changes in volume of confidence ellipsoids for $\underline{\beta}$ when changing from $\hat{\underline{\beta}}$ to $\hat{\underline{\beta}}_{(i)}$ as the vector of point estimates. If we ignore the intercept term, the logarithm of the ratio of the two volumes for $(1-\alpha)$ 100% confidence ellipsoids can be expressed in the following reduced form:

$$\begin{aligned} \log \left(\frac{\text{Vol}(E_0)}{\text{Vol}(E_{(i)})} \right) &= \frac{1}{2} \log \left(\frac{1 - h_i}{1 - 1/n} \right) \\ &+ \left(\frac{p-1}{2} \right) \log \left(\frac{n-p-1}{n-p-t_i^2} \cdot \frac{F(1-\alpha; p-1, n-p)}{F(1-\alpha; p-1, n-p-1)} \right) \end{aligned}$$

The above quantity can be both negative and positive. A large positive value indicates the presence of an outlier, while a large negative value indicates a point with extremely high leverage. Most observations should yield quantities close to zero.

The developers of these diagnostics have suggested various criteria. BKW (p. 28) consider $(1 \pm 3p/n)$ as the cutoff points for COVRATIO, with any observation which yields a value not in this interval classified as influential. Andrews and Pregibon (pp. 88-89) indicate methods by which to calculate average moments of $R_i^1(X^*)$, and to calculate percentage points for the $\min R_i^1(X^*)$. However, they encourage that their statistic be used for detection and not for assessment at a given level of significance. Cook and Weisberg merely state that you want values of their statistic near zero.

As for the previous statistics, a simple listing, a histogram, or a stem-and-leaf display is encouraged. A plot of the statistics versus the case number or the value of h_i could prove informative.

2.7 EFFECTS ON INDIVIDUAL PARAMETERS

In many instances the analyst is primarily interested in the effects of the various observations on one particular parameter. There are two measures, closely related to each other, that can be used to study this effect.

1. The first measure is due to BKW (1980, p. 13). In section 2.2 we showed that

$$\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)} = \frac{(X'X)^{-1} \underline{x}_i r_i}{(1-h_i)} .$$

If we are interested in the change in the estimate of the j th parameter, we need the j th element of the vector $\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)}$. In order to find this term, let $C = (X'X)^{-1}X'$, with c_{ij} being the ij th element. Then

$$DFBETA_{ij} = (\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})_j = \frac{c_{ji} r_i}{(1-h_i)} .$$

The authors suggest standardizing by an estimate of the variance, where $\text{Var}(\hat{\beta}_j) = \sigma^2 (X'X)^{-1}_{jj}$. But $CC' = (X'X)^{-1}$, so

$$(X'X)^{-1}_{jj} = \sum_{k=1}^n c_{jk}^2 .$$

All of this gives us a scaled measure of the change in the estimate of the j th parameter when the i th observation is deleted. This measure, expressed in reduced form, is:

$$\begin{aligned} \text{DFBETAS}_{ij} &= \frac{\hat{\beta}_j - \hat{\beta}_{(i)j}}{s(i)\sqrt{(X'X)^{-1}_{jj}}} = \frac{c_{ji}r_i}{(1-h_i)s(i)\sqrt{\sum_{k=1}^n c_{jk}^2}} \\ &= r_i^* \frac{c_{ji}}{\sqrt{(1-h_i) \sum_{k=1}^n c_{jk}^2}} \end{aligned}$$

2. The second measure is an adaptation of Cook's distance function when there is only interest in one parameter. Cook and Weisberg (1980, p. 500) define $\underline{\psi} = L\underline{\beta}$, where L is a $q \times p$ matrix of rank q . For these q different linear combinations of the parameters, they define the distance function to be

$$D_i(\underline{\psi}) = \frac{(\hat{\underline{\psi}} - \hat{\underline{\psi}}_{(i)})' [L(X'X)^{-1}L']^{-1} (\hat{\underline{\psi}} - \hat{\underline{\psi}}_{(i)})}{qs^2}$$

Let L be the elementary vector $\underline{e}_j$ with a 1 in the j th position and all other elements equal to 0. Then $\underline{\psi} = \beta_j$ and

$$\begin{aligned} D_i(\hat{\beta}_j) &= \frac{(\underline{e}_j'\hat{\beta} - \underline{e}_j'\hat{\beta}_{(i)})' [\underline{e}_j'(X'X)^{-1}\underline{e}_j]^{-1} (\underline{e}_j'\hat{\beta} - \underline{e}_j'\hat{\beta}_{(i)})}{s^2} \\ &= \frac{r_i^2 c_{ji}^2}{s^2(1-h_i)^2 \sum_{k=1}^n c_{jk}^2} = t_i^2 \left(\frac{c_{ji}^2}{(1-h_i) \sum_{k=1}^n c_{jk}^2} \right) \end{aligned}$$

The last term is the reduced form of this measure. It is extremely close to the square of DFBETAS_{ij} . The only difference

is the replacement of $s(i)$ with s , a characteristic that has differentiated the measures developed by BKW from the comparable measure developed by Cook and Cook and Weisberg.

Belsley, Kuh and Welsch (p. 28) recommend a size-adjusted cutoff of $2/\sqrt{n}$. Cook and Weisberg do not recommend any cutoff points for $D_i(\hat{\beta}_j)$. The $DFBETAS_{ij}$ and $D_i(\hat{\beta}_j)$ can be displayed in similar ways to the other diagnostics. Histograms, stem-and-leaf displays, listings, and plots versus case number are the most appropriate.

2.8 PARTIAL REGRESSION (RESIDUAL) PLOTS

Belsley, Kuh and Welsch (p. 30) advocate the use of partial regression leverage plots as a diagnostic tool. These plots and a method of development for them are explained in Mosteller and Tukey (1977, pp. 343-344). To discuss the methods of these plots, some new notation is needed. Let X_j denote the vector of n values associated with the j th carrier (independent variable). Let $\underline{r} = \underline{r}_{Y.ALL}$ be the vector of residuals from the full regression model. Let $\underline{r}_{Y.(j)}$ denote the vector of residuals when y is regressed on all the independent variables except X_j . Let $\underline{r}_{j.(j)}$ denote the vector of residuals obtained by regressing X_j on all the other independent variables. Then, the partial regression plot for β_j is defined as the plot of $\underline{r}_{Y.(j)}$ vs. $\underline{r}_{j.(j)}$. This plot has some important characteristics:

1. The least squares estimate of the slope of this plot is $\hat{\beta}_j$, the estimated coefficient for X_j in the full model, and the Y -intercept is 0.
2. The residuals from this least squares fit are the residuals associated with the full estimated regression equation.
3. The relative weight of any observation in determining the value of $\hat{\beta}_j$ is more obvious and this information has been separated from the relationships of this data point with the other independent variables.

This last characteristic is what makes partial regression plots so useful as a regression diagnostic tool. By the definition of the plot, it appears that results from three different regressions are needed. However, Velleman

and Welsh (1981) have derived the following reduced forms:

$$(r_{j \cdot (j)})_i = \frac{c_{ji}}{\sum_{k=1}^n c_{jk}^2}$$

and

$$(r_{Y \cdot (j)})_i = r_i + \hat{\beta}_j (r_{j \cdot (j)})_i$$

Therefore, knowing the residuals and coefficient from fitting the full model, and knowing the elements of the C matrix defined in section 2.7, we can retrieve the two sets of residuals necessary for any partial regression plot.

2.9 INFLUENCE ON SUBSET OF β

If the analyst is interested in a subset of the components of β , Cook and Weisberg, by using their L matrix and distance function defined in the previous section, have shown (p. 501) that the distance function can be expressed in the form

$$D_i(\psi) = \frac{t_i^2}{q} \left(\frac{h_i - \mu_i}{1 - h_i} \right) \quad (6)$$

where q is the number of components in the subset of interest, t_i and h_i are as previously defined, and μ_i is the i th diagonal element of the hat matrix associated with those columns of the X matrix concerned with the $(p-q)$ components of β that aren't of interest. Therefore, in order to calculate $D_i(\psi)$, the analyst needs the results of two separate regressions on all the observations. The calculated values for this measure would be analyzed and displayed the same as D_i .

2.10 DIAGNOSTICS FOR WEIGHTED LEAST SQUARES

In all techniques previously described, it has been assumed that the model has constant variance; i.e., $\text{Var}(y_i) = \text{Var}(\epsilon_i) = \sigma^2$, $i = 1, \dots, n$. A number of diagnostics depend on some commonly understood "average" value for their interpretation. For instance, a leverage point is defined as an

observation where $h_i > 2 \times \text{"average value"}$. An important question is what diagnostics should be used, and what modifications should be made to existing techniques when the variance structure of the random error term is not so simple. That problem will be considered in this section.

Consider the model $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, where all terms are as defined before, with the exception that $\text{Var}(\underline{\varepsilon}) = \sigma^2 D$, where D is a non-singular ($n \times n$) diagonal matrix having values d_i , $i = 1, \dots, n$ on the main diagonal. The usual way to calculate the least squares estimate of $\underline{\beta}$ is to use the process of weighted least-squares, and determine $\hat{\underline{\beta}}$ as $\hat{\underline{\beta}} = (X'WX)^{-1}X'W\underline{Y}$, where $W = D^{-1}$ and $w_i = 1/d_i$. This process however, is merely solving the set of normal equations associated with a transformed set of X and Y values. Let $U = W^{1/2}X$ and $\underline{Z} = W^{1/2}\underline{Y}$. This transformation gives $U\underline{\beta} = \underline{Z}$, or $U'U\underline{\beta} = U'\underline{Z}$, with solution $\hat{\underline{\beta}} = (U'U)^{-1}U'\underline{Z} = (X'WX)^{-1}X'W\underline{Y}$ which are the solutions to the weighted least squares problem. In terms of U and $\underline{Z}$, all points should have comparable weight. Because of this, any calculation and interpretation of regression diagnostics should be done in terms of U and $\underline{Z}$, and with the hat matrix given by $H_W = U(U'U)^{-1}U' = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$. By doing the analysis in terms of U , $\underline{Z}$ and H_W , the calculation, the interpretation, and the cutoff points for the various diagnostics will remain the same as those developed in the earlier sections.

The fact that zero covariance was assumed means there is one row of the U matrix corresponding to exactly one row of the X matrix. Thus, given any row of U that is exhibiting a large amount of influence, we can determine which row of X is, proportionate to its weight, exhibiting a large amount of influence. This ease of interpretation is not possible if a more general error covariance structure is assumed, namely $\text{Var}(\underline{\varepsilon}) = \Sigma$, where Σ is an arbitrary $n \times n$ symmetric, positive-definite matrix.

3. IMPLEMENTATION

When discussing the implementation of a regression diagnostic system, the variety of the diagnostics evaluated depends heavily on the output from the multiple regression program used, and the computer system and software available to the experimenter. In this chapter, three different levels of implementation will be discussed. The three levels will require increasing amounts of calculations. Listed below are the diagnostics that could be calculated at each level.

Level I

h_i , t_i , r_i^* , D_i , $DFFITS_i$, $Atkin-T_i$, $R_i^1(X^*)$ COVRATIO_i,
and Cook's Volume Ratio

Level II

(all of the above), $DFBETAS_{ij}$, Cook's $D_i(\hat{\beta}_j)$,
Partial Regression Plots

Level III

Cook's $D_i(\psi)$ for subset of $\underline{\beta}$

3.1 LEVEL I IMPLEMENTATION

In this section it is assumed that the experimenter has available the following information from a previously run least-squares multiple regression program (or from calculations done independent of the regression program): $(\underline{r}, \underline{h}, s, n, p)$. Most regression programs do not output $\underline{h}$, but do output $\underline{t}$, the vector of standardized residuals. If so, then $h_i = 1 - (r_i/st_i)^2$, and all the necessary information is available. The basic building blocks of all the diagnostics in this section are h_i , t_i (or r_i^*), n and p . We will not show in the implementation how to calculate $s(i)$, although the calculation based on Equation 2 of the previous chapter, is straightforward. This calculation isn't needed because the following expression can be derived:

$$r_i^* = t_i \left[\frac{n-p-1}{n-p-t_i^2} \right]^{1/2}$$

Using this relationship there is no need for $s(i)$. However, if n is small and some t_i is very large, the above formula won't be defined, and $s(i)$ will have to be calculated. Anyone interested in studying the standard deviation terms can easily recover them in the usual case, since $s(i) = (t_i/r_i^*)s$. The order in which terms are calculated is somewhat immaterial. The dependence of the various diagnostics is shown in Exhibit 1.

The numbering of the various terms in Exhibit 1 correspond to the order of the listing of the calculated diagnostics for the example of the next chapter (see Appendix C). Cook and Weisberg's volume ratio (#9) depends on table values from the F-distribution for the particular $100(1-\alpha)\%$ confidence region desired. However, if n is very large compared to p , then the ratio of the two F-values is very close to 1, and the formula for $\log(V_0/V_{(i)})$ can be correspondingly simplified.

An implementation at Level I should have the following features:

- (1) The analyst must have the capability of getting a complete listing (with case ID numbers) of one or more of the diagnostics.
- (2) The analyst must have the capability to get a stem-and-leaf display (or something similar) for each of the diagnostics.
- (3) The program must be able to accept a cutoff value for a particular diagnostic, and output the case ID and the diagnostic value for the points whose diagnostic values exceed the cutoff point. A similar feature of interest to individuals using the BKW diagnostics would be the automatic calculation of the defined cutoff values associated with these statistics, followed by the output of those cases exceeding the cutoff value(s).
- (4) The program should be capable of generating plots of each statistic versus case number (1 to n), or versus any other diagnostic.
- (5) The program should be capable of producing normal and half-normal probability plots.

EXHIBIT 1. Dependency structure for single case regression diagnostics: Implemental Level I.

The first two features are extremely important. If an analyst has a listing of the calculated values of a particular diagnostic without a visual display indicating the relative distribution of these values, the analysis is much less informative. Conversely, if the analyst has a visual display and can see the extreme values, but has no way to identify them, much information is lost. The ideal output would have all the extreme values on the visual display identified by case ID.

In designing a program to use as part of an interactive system, the user must be given some choices, but control should also be maintained. A possible design for a Level I implementation is given below. As a first measure of control, the diagnostics have been put into three overlapping groups. The three groups are:

- (1) BKW - This group contains four diagnostics possible to calculate at this level that were given in the book by Belsley, Kuh and Welsch. These are h_i , r_i^* , $DFFITS_i$, and $COVRATIO_i$.
- (2) Cook - This group contains diagnostics suggested by Cook and Cook with Weisberg. These diagnostics are t_i , D_i and Cook's Volume Ratio. The values of h_i are also included.
- (3) Atkinson - This group includes h_i , r_i^* , T_i and $R_i^1(X^*)$. The Andrews-Pregibon statistic was included in this group so that all discussed diagnostics would be included in at least one of the three groupings. This allows a user to get any single diagnostic by requesting a subset of a particular grouping.

The design is given in two parts. The first part is a pseudo-flowchart showing an overview of the possible routines that should be a part of the implementation. The second part contains a brief description of what each subroutine needs, and should be able to do.

Level I implementation of single case regression diagnostics.

Subroutine PACK

1. Input

a. Name of package chosen

(1) BKW - (Diagnostics #1, #3, #5, #8)

(2) Cook - (Diagnostics #1, #2, #4, #9)

(3) Atkinson - (Diagnostics #1, #3, #6, #7)

2. Determine

a. Does user want all or subset of diagnostics in package.

b. If all, set numbers in package equal to values of diagnostics required.

c. If subset, individually query which diagnostics in package are desired.

3. Output

a. If all, output the four numbers associated with the particular package.

b. If subset, output numbers associated with subset desired.

c. k, the number of diagnostics requested.

Subroutine CALC

1. Input
 - a. $\hat{h}$, a vector of "hat" values
 - b. s , least squares standard deviation
 - c. n - sample size
 - d. p - number of parameters in model
 - e. The numbers of the diagnostics desired (i.e., #1, #2, #4, #9)
(Output from subroutine PACK)
2. Compute
 - a. $\underline{t}_i$ (or the vector $\underline{t}$)
(Compute $\underline{h}$, if $\underline{t}$ given.)
 - b. Compute the vector of values for each of the diagnostics requested (using formulas in Exhibit 1).
(If #1 or #2 requested, strictly map the values into their locations.)
 - c. If #9 is desired, query for F values.
3. Output
 - a. k vectors, where k is the number of diagnostics requested.
(Let $\underline{d}$ represent a vector of diagnostics. If we requested diagnostics #2, #4, #9, we would output $(\underline{d}_{(1)}, \underline{d}_{(2)}, \underline{d}_{(3)}) = (\underline{d}_2, \underline{d}_4, \underline{d}_9)$. If $k = 4$, $\underline{d}_1$ is always outputted.)

Subroutine LIST

1. Input

- a. k , where k is the number of diagnostics requested,
- b. The vector of case IDs, containing an identifying number for each observation.
- c. $\underline{d}_{(1)}, \underline{d}_{(2)}, \dots, \underline{d}_{(k)}$, the k vectors of diagnostics calculated in subroutine CALC.

2. Compute

- a. Determine if all the diagnostics whose calculations were requested are wanted in LIST. If not, determine which are wanted.
- b. Determine if LIST is to be printed after a sort is done on a particular diagnostic. If yes, determine diagnostic and conduct sort.

3. Output

- a. Print out case number and diagnostics requested in this subroutine for each of the n cases. (Nothing to pass)

Subroutine STEM

1. Input
 - a. k , number of diagnostics requested in subroutine PACK
 - b. $\underline{d}_{(1)}, \underline{d}_{(2)}, \dots, \underline{d}_{(k)}$; the k vectors of diagnostics calculated in subroutine CALC
2. Compute
 - a. Determine which diagnostics, out of k originally selected, the analyst wishes to display.
 - b. For each diagnostic, go through a routine to tabulate a stem-and-leaf display. (A desirable but difficult feature would be to give case number of a small fraction of observations at each end of display.)
3. Output
 - a. Output display of each of the stem-and-leaf displays desired. (Nothing to pass)

Subroutine CRIT

1. Input

- a. p - number of parameters in model
- b. n - number of observations
- c. k - number of diagnostics requested
- d. $(1), \dots, (k)$, the number of diagnostics requested
- e. $\underline{d}_{(1)}, \underline{d}_{(2)}, \dots, \underline{d}_{(k)}$. The vectors of calculated diagnostics.
- f. Case ID for each observation

2. Compute

- a. For each diagnostic, query user if cutoff value(s) is (are) desired, and on which side the extreme points fall.
- b. If possible, give user the flexibility of giving cutoff value in terms of p and n .
- c. For each diagnostic, determine values of diagnostic that exceed cutoff point, and carry along case ID number.
- d. If no cutoff value is given, ignore diagnostic.

3. Output

- a. Print out listing (for each diagnostic) of case ID and diagnostic value for all observations exceeding cutoff.
(Nothing to pass back)

Subroutine PLOT

1. Input
 - a. k , number of diagnostics requested
 - b. $\underline{d}_{(1)}, \underline{d}_{(2)}, \dots, \underline{d}_{(k)}$, the vectors of diagnostics calculated
2. Compute
 - a. Determine plots wanted
 - (1) $\underline{d}_{(i)}$ versus I , an indicator going from 1 to n , or
 - (2) $\underline{d}_{(i)}$ versus $\underline{d}_{(j)}$
 - b. Allow possibility of doing more than one plot while in subroutine
3. Output
 - a. Print (or display) each of the desired plots.

Subroutine NORM

1. Input

- a. Numbers of the diagnostics calculated
- b. $\underline{d}_{(1)}, \dots, \underline{d}_{(k)}$
- c. n

2. Compute

- a. This technique should only be used with diagnostics #2 or #3. Therefore, check which of these two numbers, if any, are included in Input a. and proceed with that diagnostic.

- b. Order diagnostic from smallest to largest

- c. For $I = 1, n$

$$y(i) = (I - 0.375)/(n + 0.25) \quad (*)$$

For $I = 1, n$

$$x(I) = \Phi^{-1}(y(I)) \quad (**)$$

where Φ^{-1} is the inverse of the standard normal distribution.

3. Output

- a. Plot ordered t_i (#2) or ordered r_i^* (#3) (vertical axis) versus $x(I)$.

*Due to Blom (1958, p. 71). Other possibilities are given by Draper and Smith (1981, p. 178) and by Filliben (1975, p. 116).

**For approximations see Abramowitz and Stegun (1964, p. 933) or Ryan and Joiner (p. 4).

Subroutine HNORM

1. Input

- a. Numbers of the diagnostics calculated
- b. $\underline{d}_{(1)}, \dots, \underline{d}_{(k)}$
- c. n

2. Compute

- a. Determine which diagnostic analyst wants to do half-normal plot with (should only be #2, #3, or #6). (Could be both #3 and #6, so should allow for more than one plot.)
- b. If t_i (#2) or r_i^* (#3), take absolute value of calculated diagnostics.
- c. Order diagnostic (after absolute values) from smallest to largest.
- d. For $I = 1, n$, generate

$$y(I) = 0.5 + 0.5(I - 0.375)/(n + 0.25) \quad (*)$$

- e. For $I = 1, n$

$$x(I) = \Phi^{-1}(Y(I)) \quad (**)$$

3. Output

- a. Plot ordered diagnostic (vertical axis) versus $x(I)$.

*Other possibility, Draper and Smith (p. 180).

**For approximations see Abramowitz and Stegun (1964, p. 933) or Ryan and Joiner (p. 4).

3.2 LEVEL II IMPLEMENTATION

At Level II implementation, the set of possible diagnostics includes all those listed at Level I, plus diagnostics concerned with the effects of single cases on individual parameters. These diagnostics are DFBETAS_{ij}, Cook's D_i for individual parameters, and partial regression plots. In order to complete the calculation of these diagnostics, any implemented program must be supplied with more information. In addition to the input at Level I, Level II implementation requires (a) (X'X) and X', or (b) (X'X)⁻¹ and X', or (c) (X'X)⁻¹X'. Also, for partial regression plots, $\hat{\beta}$ is needed. The actual additional information needed for the first part is (c) C = (X'X)⁻¹X'. However, (a) or (b) could be the forms of input, with both forms requiring additional computation before they're in usable form.

A Level II program would be an extension of Level I. If #10 represents DFBETAS_{ij}, #11 represents Cook's D_i for individual parameters, and #12 represents partial regression plots, #10 could be added to the BKW group, #11 could be added to the Cook group, and #12 added to all groups (see Exhibit 1). From Chapter 2, we have

$$\text{DFBETAS}_{ij} = \frac{r_i^* c_{ji}}{\sqrt{(1-h_i) \sum_{j=1}^n c_{jk}^2}}$$

$$D_i(\hat{\beta}_j) = \frac{t_i^2 c_{ji}^2}{(1-h_i) \sum_{k=1}^n c_{jk}^2}$$

Therefore, the actual calculation priority would be #2 → #11 and #3 → #10. The coding for these two calculations could be included in subroutine CALC.

It should be noted that, in addition to requiring storage for input, and requiring more calculations, the calculation of DFFITS or D_i(β_j) requires much more storage for results. If information on β₀ is not wanted, there will be n(p-1) results generated for DFFITS_{ij} (or D_i(β_j)), j = 2, ..., p. For

this reason, a program might be designed that would require that calculation of these diagnostics be requested explicitly.

Because of its unique nature, the partial regression plot could be put into a separate subroutine. This subroutine would require the C matrix, the vector of residuals $\underline{r}$, and the vector of estimated coefficients $\hat{\underline{\beta}}$. This routine, requiring the user to choose only one independent variable at a time, would then use the equations of section 2.9, namely

$$(r_{j \cdot (j)})_i = \frac{c_{ji}}{\sum_{k=1}^n c_{jk}^2}$$

and

$$(r_{Y \cdot (j)})_i = r_i + \hat{\beta}_j (r_{j \cdot (j)})_i$$

These calculations would be followed by a plot of $(r_{Y \cdot (j)})_i$ (vertical) versus $(r_{j \cdot (j)})_i$.

3.3 LEVEL III IMPLEMENTATION

At this level we require the ability to do the following:

- (1) Run a regression model with p parameters in $\underline{\beta}$, and store the diagnostics t_i (or t_i^2) and h_i .
- (2) Run a subset of this model with (p-q) parameters, and store the diagnostics μ_i (diagonal elements of hat matrix for reduced model).
- (3) Calculate

$$D_i(\psi) = \frac{t_i^2}{q} \left[\frac{h_i - \mu_i}{1 - h_i} \right]$$

This implementation doesn't require many of the subroutines listed for Level I. It does require the ability to access the results from the two different regression runs, and put them in the same work space. The PLOT and CRIT subroutines should also be accessible.

Cook and Weisberg (1981, p. 502) mention an easier way to solve this problem if the QR factorization of X is being used. All of the formulas in Chapter 2 were expressed in the usual regression notation. However, many computer programs solve the least-squares problem by more stable numerical methods such as QR decomposition. Nearly all terms involved in regression diagnostics can be defined in terms of the results of the QR decomposition. This topic is discussed in Appendix G.

4. ANALYSIS OF EXAMPLE DATA

The data studied in this chapter consists of air pollution, socioeconomic, and mortality rate values for 112 different SMSAs (Standard Metropolitan Statistical Areas) for the year 1969. This data was originally analyzed in the book by Lave and Seskin (1977). For this example 11 different independent variables were used, with unadjusted total mortality rate as the dependent variable. The order, program name, and definition of the variables are:

X_1	Min-S:	Smallest biweekly sulfate reading ($\mu\text{g}/\text{m}^3 \times 10$)
X_2	Mean-S:	Arithmetic mean of the biweekly sulfate readings ($\mu\text{g}/\text{m}^3 \times 10$)
X_3	Max-S:	Largest biweekly sulfate reading ($\mu\text{g}/\text{m}^3 \times 100$)
X_4	Min-P:	Smallest biweekly suspended particulate reading ($\mu\text{g}/\text{m}^3 \times 10$)
X_5	Mean-P:	Arithmetic mean of biweekly suspended particulate readings ($\mu\text{g}/\text{m}^3$)
X_6	Max-P:	Largest biweekly suspended particulate reading ($\mu\text{g}/\text{m}^3$)
X_7	Pop-Dens:	SMSA population density (per square mile $\times 0.1$)
X_8	GTE65:	Percent of SMSA population at least 65 years old ($\times 10$)
X_9	Non-Wh:	Percent of nonwhites in SMSA population ($\times 10$)
X_{10}	Poor:	Percent of SMSA families with income below the poverty level ($\times 10$)
X_{11}	Log-Pop:	The logarithm of SMSA population ($\times 100$)
Y	Mort-RT:	Unadjusted total mortality rate (per 100,000).

A listing of the observations is given in Appendix A. The first column is the ID for the particular SMSA. A list of these numbers with the name of each area is given in Appendix B.

All of the calculations, displays, and plots were done using the Biometrics Center PDP 11/70, operated by Pacific Northwest Laboratory. The Minitab 80 statistical program was used exclusively. In the analysis it is assumed that the unadjusted total mortality rate (Y) is a linear function of the 11 independent variables as defined above. It is also assumed that the error term has constant variance across the 112 different SMSAs. The fitted model and some pertinent statistics are given below. After discussing the fit, the various diagnostics will be considered in the same order as they were discussed in Chapter 2.

4.1 THE FITTED MODEL

The Minitab output for the least squares regression of unadjusted total mortality rate on the 11 independent variables is given in Figure 1. The results show an R^2 greater than 80%, a standard deviation about the regression line of 66.5 (compared to $s_y = 147.7$), and the t-ratios of 7 of the 11 slope coefficients larger than 1.8. These statistics indicate that the regression equation is worthwhile, but they don't indicate anything about whether or not there are extreme points in the sense of having too much influence or having residuals that are too large. (A complete Minitab regression output does flag observations with large t_i .) The rest of the chapter will consist of analysis of the various diagnostics in an attempt to determine if there are any extreme points, and if so, the identity of these extreme points.

4.2 ANALYSIS OF HAT MATRIX DIAGONALS (See Section 2.1)

The h_i values were calculated (based on the model with a y-intercept term and 11 independent variables) for each the 112 SMSAs. The complete listing of these values, along with the values of many other diagnostics is given in Appendix C. A stem-and-leaf display for the values of h_i is given below in Figure 2.

The display shows that there are definitely some extreme points in terms of the h_i values. With $p = 12$ and $n = 112$, the average value of h_i is 0.107. The BKW suggested cutoff point is $(2p/n) = 0.214$. The stem-and-leaf display indicates that 9 or 10 data points exceed the cutoff value with four values exceeding $3p/n$. The 10 SMSAs that are defined as leverage points ($h_i > 0.214$) are listed in Table 1.

THE REGRESSION EQUATION IS

$$Y = 387. -0.0384 X_1 + 0.633 X_2 -0.0527 X_3 + 0.434 X_4 +0.0555 X_5 + 0.130 X_6 + 0.130 X_7 + 6.40 X_8 + 0.204 X_9 + 0.511 X_{10} - 0.428 X_{11}.$$

	COLUMN	COEFFICIENT	ST. DEV. OF COEF.	T-RATIO = COEF/S.D.
	--	387.0117	115.0568	3.36
X1	MIN-S	-0.0384	0.5265	-0.07
X2	MEAN-S	0.6329	0.3500	1.81
X3	MAX-S	-0.0527	0.0784	-0.67
X4	MIN-P	0.4343	0.6262	0.69
X5	MEAN-P	0.0555	0.4219	0.13
X6	MAX-P	0.1296	0.0707	1.83
X7	POP-DENS	0.1304	0.0520	2.51
X8	GTE65	6.4030	0.3742	17.11
X9	NON-WH	0.2037	0.0908	2.24
X*	POOR	0.5113	0.2406	2.13
X*	LOG-POP	-0.4277	0.1923	-2.22

THE ST. DEV. OF Y ABOUT REGRESSION LINE IS

S = 66.53

WITH (112-12) = 100 DEGREES OF FREEDOM

R-SQUARED = 81.7 PERCENT

R-SQUARED = 79.7 PERCENT, ADJUSTED FOR D.F.

ANALYSIS OF VARIANCE

DUE TO	DF	SS	MS=SS/DF
REGRESSION	11	1980316	180029
RESIDUAL	100	442606	4426
TOTAL	111	2422922	

FIGURE 1. Regression output for example.

STEM-AND-LEAF DISPLAY OF H(1)
 LEAF DIGIT UNIT = 0.0100
 1 2 REPRESENTS 0.12

```

22 +0* 2233333344444444444444
(48) +0. 55555556666666677777777778888888889999999999
42 1* 000000000111111123333334
16 1. 555667
10 2* 1233
6 2. 57
4 3*
4 3. 67
2 4* 2
1 4.
1 5*
1 5.
1 6*
1 6.
1 7*
1 7.
1 8*
1 8. 5

```

FIGURE 2. Stem-and-leaf display for h_i .

TABLE 1. Leverage points (cutoff = 0.214)

<u>SMSA Code #</u>		<u>Value of h_i</u>
499	Jersey City, NJ	0.853
504	Buffalo, NY	0.428
434	Tampa, FL	0.373
441	Honolulu, HI	0.362
443	Chicago, IL	0.272
583	Huntington, WV	0.254
440	Savannah, GA	0.238
510	New York, NY	0.232
582	Charleston, WV	0.229
431	Jacksonville, FL	0.218

As mentioned in Chapter 2, the designation of a point as a leverage point does not indicate the point has high influence in determining our regression equation. It does, however, flag these points as potential influence points, and indicates that these SMSAs have values near the edge of the independent variable hull (IVH). For example, Jersey City has an h_i value more than twice that of any other SMSA. The size of this value is caused principally by the values of the independent variable Pop-Dens. The population density of Jersey City is nearly three times larger than the next highest value (New York), and is seven times larger than all but two values. In lieu of this fact, a logarithmic transformation of the population density is probably warranted, but it was not done for this example.

4.3 ANALYSIS OF "STANDARDIZED" RESIDUALS (See Section 2.4)

The values of the standardized residuals (t_i) and the "studentized" residuals (r_i^*) were calculated for all SMSAs and have a complete listing in Appendix C. Stem-and-leaf displays are given below in Figures 3A and 3B. As discussed earlier, the only difference in the two diagnostics is in the scaling factor involved. For residuals far removed from zero, the value of $|r_i^*|$ will be larger than the value of $|t_i|$. This causes the stem-and-leaf display for r_i^* to be spread slightly more than the display of t_i . However, the relative rankings of the SMSAs with respect to the magnitude of each of these diagnostics will be the same, since $|t_i| > |t_j|$ implies $|r_i^*| > |r_j^*|$. Therefore, both diagnostics will identify the same extreme points.

As mentioned in section 2.4, Atkinson has recommended studying r_i^* in a half-normal plot. This plot appears as Figure 4. Whereas the stem-and-leaf display for r_i^* appears to have from three to six extreme points, the half-normal plot clearly identifies four values as falling outside the hypothesized pattern.

A usual cutoff value for extreme points of "standardized residuals" is a magnitude of 2. For the Lave and Seskin 1969 data, the extreme points are listed in Table 2. Both diagnostics identify exactly six observations as extreme points.

STEM-AND-LEAF DISPLAY OF T(1)
 LEAF DIGIT UNIT = 0.1000
 1 2 REPRESENTS 1.2

```

1  -3. 8
2  -* 4
2  -2.
3  -* 1
5  -1. 88
15 -* 4333321100
31 -0. 9888888776665555
(26) -0* 44444333332222222211000000
55 +0* 000001111122222334444
34 +0. 55555566777778899999
14 1* 0000233444
4 1. 5
3 2* 1
2 2. 7
1 3* 2

```

FIGURE 3A. Stem-and-leaf display for standardized residuals, t_i .

STEM-AND-LEAF DISPLAY OF R(1)*
 LEAF DIGIT UNIT = 0.1000
 1 2 REPRESENTS 1.2

```

1  -4. 1
2  -3. 5
2  -3.
2  -2.
3  -2. 2
5  -1. 88
15 -1. 4333321100
31 -0. 9888888776665555
(26) -0* 44444333332222222211000000
55 +0* 000001111122222334444
34 +0. 55555566777778899999
14 1* 0000233444
4 1. 5
3 2* 1
2 2. 8
1 3* 4

```

FIGURE 3B. Stem-and-leaf display for studentized residuals, r_i^* .

FIGURE 4. Half-normal plot for $|r_i^*|$.

TABLE 2. Extreme values for "standardized" residuals (cutoff = ± 2).

<u>SMSA Code #</u>	<u>Location</u>	<u>t_i</u>	<u>r_i^*</u>
434	Tampa, FL	-3.84	-4.14
441	Honolulu, HI	-3.40	-3.60
484	Duluth, MN	3.29	3.46
548	Wilkes-Barre, PA	2.78	2.88
692	Fargo, ND	-2.19	-2.24
403	Montgomery, AL	2.10	2.14

At this point let us compare the extreme points identified by the two types of diagnostics already discussed, namely h_i and t_i or r_i^* . The two locations with the largest h_i values (Jersey City and Buffalo) do not appear in Table 2. In fact, the value of t_i is -0.05 for Jersey City and -0.27 for Buffalo. However, Tampa and Honolulu, which had values of h_i exceeding $3p/n$, have the two largest values of the standardized residuals. The other four SMSAs listed in Table 2 had much smaller values of h_i , with Montgomery (0.15) being the only one with a value greater than the mean of the h_i (0.107). This comparison points out that a correct analysis of the results of a regression fit must consider at least two different types of points, those points that are extreme in terms of their own values on the independent variables, and those observations that are true outliers (in y) based on the assumed model.

4.4 ANALYSIS OF EFFECT ON $\hat{Y}_i$ (See Section 2.5)

There are three diagnostics that measure the effect of an individual observation on the regression estimates of $\underline{Y}$. As previously defined, these are Cook's D_i , the DFFITS $_i$ of Belsley, Kuh and Welsch, and Atkinson's T_i . The complete listing of the calculated values of these diagnostics is given in Appendix C. Figures 5A, 5B and 5C below show stem-and-leaf displays for DFFITS, D_i and T_i , respectively.

STEM-AND-LEAF DISPLAY OF DFFIT
LEAF DIGIT UNIT = 0.1000
1 2 REPRESENTS 1.2

```

1      -3* 1
2      -2. 7
2      -2*
2      -1*
2      -1*
7      -0. 76655
(50)   -0* 44443333222222222222222111111111000000000000000000000
.55     +0* 0000000000000000000111111111122222222222223333344
6       +0. 55799
1       1* 0

```

FIGURE 5A. Stem-and-leaf display for DFFITS.

```
STEM-AND-LEAF DISPLAY OF D(1)
LEAF DIGIT UNIT = 0.0100
1 2 REPRESENTS 0.12
```

```
(107) +0* 00000000000000000000000000000000000000000000000...  
      5 +0. 667  
      2 1*  
      2 1.  
      2 2*  
      2 2.  
      2 3*  
      2 3.  
      2 4*  
      2 4.  
      2 5* 4  
      1 5.  
      1 6*  
      1 6.  
      1 7* 3
```

FIGURE 5B. Stem-and-leaf display for D_j .

STEM-AND-LEAF DISPLAY OF ATKIN-T
 LEAF DIGIT UNIT = 0.1000
 1 2 REPRESENTS 1.2

```

52 +0* 00000000001111111111111122222222222222233333334444444
(38) +0. 55555555666666666777777777778888889999
22 1* 011222233
13 1. 666778
7 2* 12
5 2. 679
2 3*
2 3.
2 4*
2 4.
2 5*
2 5.
2 6*
2 6.
2 7*
2 7. 8
1 8*
1 8.
1 9* 2

```

FIGURE 5C. Stem-and-leaf display for T_i .

In comparing the three displays it is obvious that two observations are very extreme for each of the diagnostics. A half-normal plot of the T_i also indicates two extreme values (see Figure 6). It turns out that these are the same two observations. There is some difference in the displays, however. D_i and T_i separate the two most extreme points from the rest of the points much more than DFFITS does. Also DFFITS gives a direction to the change, while the other two diagnostics don't. In choosing between the three diagnostics, the analyst will undoubtedly want to take into account the justification for each of the diagnostics. However, the greater separation of the last two diagnostics appears to be a decided advantage.

Whichever diagnostic is chosen, the rankings of the observations in magnitude of the calculated values of the diagnostic is the same. This result, similar to that for the two types of standardized residuals, is readily apparent through analyzing Equations 3, 4 and 5. In

FIGURE 6. Half-normal plot of T_i .

choosing points of influence with respect to these diagnostics, Table 3 lists all points whose $DFFITS_i$ value exceeds the cutoff suggested by BKW, namely $2\sqrt{p/n} = 0.655$.

TABLE 3. Points of influence for $DFFITS_i$, D_i , and T_i
(cutoff value: $|DFFITS_i| > 0.655$).

<u>SMSA #</u>	<u>Location</u>	<u>$DFFITS_i$</u>	<u>D_i</u>	<u>T_i</u>
434	Tampa, FL	-3.19	0.732	9.22
441	Honolulu, HI	-2.71	0.545	7.82
484	Duluth, MN	1.01	0.076	2.90
548	Wilkes-Barre, PA	0.94	0.069	2.72
403	Montgomery, AL	0.90	0.065	2.60
440	Savannah, GA	0.77	0.049	2.23
432	Miami, FL	-0.74	0.045	2.14

The ranking of the first four extreme points is the same ranking as that for the standardized residuals (Table 2). The separation with these diagnostics is much greater between the first two SMSAs and the second two SMSAs. Although the BKW criteria chose seven locations as points of influence, only the first two points appear as extreme when considered against the diagnostic values for the remainder of the observations. In this regard the cutoff criteria suggested by BKW appears to be excessively liberal. For example, Miami has $h_i = 0.136$ and $t_i = 1.85$, neither of which is exceedingly large. However, the $DFFITS$ term, which is a combination of these two terms, falls above the cutoff point. Instead of using an arbitrary cutoff point, it is suggested that the analyst look closely at the separation of the extreme point diagnostic values from the remainder of the values.

4.5 DIAGNOSTICS BASED ON DETERMINANTS (See Section 2.6)

As discussed in Chapter 2, there are three diagnostic measures based on the determinants of various matrices. These diagnostics are $R_i^1(X^*)$, $COVRATIO_i$, and Cook-Weisberg's Volume ratio. As with previously discussed diagnostics, a listing of the calculated values for these diagnostics is in Appendix C. For the volume ratio diagnostic, a 95% confidence level was used. Although the three diagnostics are all based on determinants, as a group they are more

dissimilar than the last two groups. Because of this, each diagnostic will be considered separately.

First consider $R_i^1(X^*)$, suggested by Andrews and Pregibon. The stem-and-leaf display is given in Figure 7. Extreme values for this diagnostic are values much less than 1, making interpretation of this diagnostic less natural. The display shows four values much removed from the others. The ID and location corresponding to these values is given in Table 4.

The Andrews-Pregibon statistic flagged the four cases with the largest values of h_i , indicating a sensitivity to leverage points. This turns out to be a characteristic of the three methods based on determinants. In fact, the observations with the 10 lowest values of $R_i^1(X^*)$ also have the 10 highest values of h_i , with the only differences in ranking occurring in spots 2 and 3, where Tampa and Honolulu moved ahead of Buffalo (second largest value of h_i). All of the other diagnostic procedures exhibited more sensitivity to the size of the residuals.

The second diagnostic is the COVRATIO_i proposed by Belsley, Kuh and Welsch. The stem-and-leaf display is given in Figure 8. Extreme points for this diagnostic occur much below 1, or much above 1. BKW suggest that a large value will occur when there is a point with high leverage but small residual, while a small value occurs when there is a large residual. The developers give cutoff values of $(1 - 3p/n, 1 + 3p/n) = (0.68, 1.32)$. These cutoff points yield 12 different points. Analyzing the spread of the values, there appear to be six values that could be considered extreme. The IDs and locations of the six extreme points are given in Table 5.

As expected, the large values occur for the two SMSAs that had highest values of h_i (and small residuals), while the small values occur for the four observations with the largest standardized residuals. A surprise is that Duluth is considered more extreme than Honolulu (the SMSA with second largest residual). The fact that h_i was large for Honolulu must be bringing the value back toward 1. This interpretation, if true, doesn't appear to be a property that you want in a diagnostic intended to identify leverage points and points having large residuals.

STEM-AND-LEAF DISPLAY OF AN-PREG
 LEAF DIGIT UNIT = 0.0100
 1 2 REPRESENTS 0.12

```

1      1* 4
1      1.
1      2*
1      2.
1      3*
1      3.
1      4*
1      4.
2      5* 3
4      5. 67
4      6*
4      6.
7      7* 234
10     7. 668
20     8* 1122333344
51     8. 55666677888888888888889999999999
(48)   9* 000000001111111111222222333333333344444444444444
13     9. 5555555666667

```

FIGURE 7. Stem-and-leaf display of Andrews-Pregibon $R_i^1(X^*)$.

TABLE 4. Extreme values for Andrews-Pregibon $R_i^1(X^*)$.

<u>SMSA#</u>	<u>Location</u>	<u>$R_i^1(X^*)$</u>
499	Jersey City, NJ	0.147
434	Tampa, FL	0.534
441	Honolulu, HI	0.564
504	Buffalo, NY	0.571

STEM-AND-LEAF DISPLAY OF COVRATIO

LEAF DIGIT UNIT = 0.1000

1 2 REPRESENTS 1.2

[illegible]

FIGURE 8. Stem-and-leaf display for COVRATIO_j.

TABLE 5. Extreme values for COVRATIO.

<u>SMSA #</u>	<u>Location</u>	<u>COVRATIO</u>
499	Jersey City, NJ	7.65
504	Buffalo, NY	1.95
434	Tampa, FL	0.26
484	Duluth, MN	0.31
441	Honolulu, HI	0.40
548	Wilkes-Barre, PA	0.48

The last diagnostic in this group is the volume ratio of two confidence regions proposed by Cook and Weisberg. As mentioned before, the 95% confidence level was chosen. A stem-and-leaf display is shown in Figure 9. Extreme points for this diagnostic are those having large positive or large negative values. A large positive value lends support to the presence of an outlier, while a large negative value indicates a point with very high leverage. In this aspect, Cook's volume ratio detects and separates "influence" points into two categories, as did COVRATIO.

A listing of the extreme points for this diagnostic is given in Table 6. This diagnostic picked the same six locations as extreme points that were chosen by COVRATIO. The one advantage is that this diagnostic is centered at zero, while COVRATIO is centered at 1. This property makes it easier to compare values on different sides of zero.

Two of the diagnostic measures discussed in this group do a good job of discriminating between the two types of extreme points, and at least one should be used as a set of complete regression diagnostics. The Andrews-Pregibon diagnostic, however, appears to offer little additional information.

4.6 EFFECTS ON INDIVIDUAL PARAMETERS (See Section 2.7)

In many uses of regression, much importance is placed on the interpretation of the values of the individual coefficients. In this section, diagnostics which measure the effect of deletion of the i th observation on the j th coefficient are studied. The y-intercept parameter will not be considered in this example, although there is no reason why in general it cannot be. The two main diagnostics used in the study of effects of observations on individual parameters are $DFBETAS_{ij}$ and Cook's $D_i(\hat{\beta}_j)$. A listing of the calculated values of $DFBETAS$ is given in Appendix D, while a listing of Cook's $D_i(\hat{\beta}_j)$ is given in Appendix E.

Although a complete analysis should include the study of these diagnostics for all of the parameters, we will only show plots for three of the parameters, while giving some summary measures on all parameters for observations that exceed cutoff values. The three parameters will be β_2 (associated with Mean-S), β_7 (associated with Pop-Dens), and β_8 (associated

[illegible]

TABLE 6. Extreme values for Cook's volume ratio.

<u>SMSA #</u>	<u>Location</u>	<u>Cook Volume Ratio</u>
499	Jersey City, NJ	-1.01
434	Tampa, FL	0.59
484	Duluth, MN	0.54
441	Honolulu, HI	0.40
548	Wilkes-Barre, PA	0.34
504	Buffalo, NY	-0.32

with GTE 65). The statistic $\hat{\beta}_2$ was the most significant estimate dealing with a pollution variable, while $\hat{\beta}_7$ and $\hat{\beta}_8$ are associated with the most significant variables in the fitted model.

Belsley, Kuh and Welsch suggest cutoff values of $\pm 2/\sqrt{n}$ for the statistic DFBETAS. If these cutoff values are used, there are a total of 78 different instances where the cutoff values are exceeded, which seems excessive. If the cutoff value is increased to $\pm 3/\sqrt{n}$, the number of instances decreases to 34. Table 7 summarizes the number of times these two cutoff values are exceeded for each of the 11 variables. Table 8 lists those SMSAs where DFBETAS exceeds the $2/\sqrt{n}$ cutoff for at least four different parameters. The value in parentheses corresponds to the number of times $3/\sqrt{n}$ is exceeded. Using the cutoff value of $2/\sqrt{n}$, these eight locations account for 58% (45/78) of the situations where the cutoff value is exceeded. Using $3/\sqrt{n}$ they account for 82% of the situations where it is exceeded. With the exception of Chicago, the locations in Table 8 are the locations where the DFFITS cutoff value was exceeded (see Table 3). This result lends support to the belief that these SMSAs are definitely the most influential. It should be noted that either Tampa (5 times) or Honolulu (3 times) have the largest absolute value of DFBETAS for 8 of the 11 parameters. As in the cases of t_i with r_i^* , and D_i with DFFITS, there will be no difference in the rankings of the observations according to the magnitude of the two diagnostic DFBETAS_{ij} and $D_i(\hat{\beta}_j)$. These diagnostics differ by a scaling factor and a square. Therefore any comparisons made using DFBETAS would yield the same results using $D_i(\hat{\beta}_j)$.

In analyzing the effects of an individual parameter, the results could be shown using a stem-and-leaf display, a histogram, or a plot. For purpose of variety, we will use plots of the diagnostics vs. the observational number (1-112). Figures 10A and 10B show the plots of the diagnostics for $\hat{\beta}_2$. The points identified on the plots are those locations where $|\text{DFBETAS}| > 2/\sqrt{n}$. In comparing the two plots, it is obvious that the better separation occurs using Cook's diagnostic while DFBETAS carries along the directional change for the parameter. A negative value means the estimated value of the parameter will increase when the i th value is deleted. A diagnostic with a more natural interpretation would be $-\text{DFBETAS}$. For this

TABLE 7. Summary of occurrences of DFBETAS exceeding cutoff values.

<u>Variable</u>	<u>Number Exceeding $2/\sqrt{n}$</u>	<u>Number Exceeding $3/\sqrt{n}$</u>
Min-S	10	3
Mean-S	7	5
Max-S	10	5
Min-P	8	5
Mean-P	8	2
Max-P	8	1
Pop-Dens	1	1
GTE 65	7	5
Non-Wh	6	3
Poor	8	2
Log-Pop	5	2

TABLE 8. Listing of SMSAs and number of parameters where DFBETAS exceeds $2/\sqrt{n}$.

<u>SMSA</u>	<u>Number of Parameter</u>	
Tampa	1,2,4,5,6,7,8,9,11	(6)
Honolulu	1,4,8,9,10,11	(5)
Montgomery	4,5,6,8,9,10	(4)
Savannah	1,2,5,6,9,11	(0)
Miami	1,2,3,4,8	(3)
Wilkes-Barre	k,2,3,4,8	(4)
Duluth	1,2,3,8	(4)
Chicago	1,4,5,6	(2)

diagnostic, the sign indicates direction of change when an observation is deleted. Both plots show the parameter estimate is affected by a number of different observations, with Duluth causing the most change. In the section on partial regression plots we will get a pictorial reason why these SMSAs cause the largest change in the estimate for $\hat{\beta}_2$.

Two other sets of plots are also given. Figures 11A and 11B plot the diagnostics for $\hat{\beta}_7$, while Figures 12A and 12B plot the diagnostics for $\hat{\beta}_8$. At first glance, the two sets are very similar, with one location dominating all others. However, the scale on the left hand side of the plots indicates how different the two cases are. For $\hat{\beta}_7$, Tampa is the only point where the cutoff value is exceeded, while for $\hat{\beta}_8$, the cutoff value is exceeded by seven points. However, the value for Tampa is so large, it overwhelms all other values. Since the absence of Tampa causes a change of approximately three standard deviations in the most significant model coefficient, it is indeed an influential point. The four non-significant values are labeled on the $\hat{\beta}_7$ plots in anticipation of the partial regression plots.

A more intuitive approach to the analysis of the effect of an individual observation on a particular coefficient is to consider the percent change in the coefficient when the observation is dropped from the list of points, i.e., $(\hat{\beta}_j(i) - \hat{\beta}_j)/\hat{\beta}_j$. These terms can be easily calculated as $c_{ji}y_i/\hat{\beta}_j$. Figures 10C, 11C and 12C show these percent changes for $\hat{\beta}_2$, $\hat{\beta}_7$ and $\hat{\beta}_8$, respectively, plotted against the observational number. Comparing these figures with their counterpart plots for $D_i(\hat{\beta}_j)$ and DFBETAS show the same SMSAs giving extreme values but the percentage figure does give a more straightforward interpretation of an observation's effect. The danger in this procedure is in comparing effects on one parameter estimate to those effects on another parameter estimate. It is easy to get a large percentage change in a coefficient that is extremely non-significant, but in most cases this is not as important as a medium change in a very important coefficient. If we consider the strength of the extreme points in the various plots, it is obvious that a large number of observations are causing a change of at least 15% in $\hat{\beta}_2$, the least significant of our three coefficients. However, for $\hat{\beta}_8$, the most significant coefficient in our model, only Tampa causes a

FIGURE 11A. Plot of $D_i(\hat{\beta}_7)$ vs. observational number.

FIGURE 11B. Plot of $DFBETAS_{i7}$ vs. observational number.

FIGURE 12A. Plot of $D_i(\hat{\beta}_8)$ vs. observational number.

FIGURE 12B. Plot of $DFBETAS_{i8}$ vs. observational number.

FIGURE 10C. Plot of percent change in $\hat{\beta}_2$ vs. observational number.

FIGURE 11C. Plot of percent change in $\hat{\beta}_7$ vs. observational number.

FIGURE 12C. Plot of percent change in $\hat{\beta}_8$ vs. observational number.

change greater than 5%. This result occurs even though there were the same number of extreme observations for both estimates using DFBETAS cutoff values.

4.7 PARTIAL REGRESSION PLOTS (See Section 2.8)

Partial regression plots can show a number of things. Among these are the following: the strength of the linear relationship between the dependent variable and a particular independent variable after the information contained in the other dependent variables have been removed from both; and, the location and influence attributable to certain observations after information in the other variables has been removed. Figures 13, 14 and 15 give the partial regression plots associated with $\hat{\beta}_2$, $\hat{\beta}_7$ and $\hat{\beta}_8$. The points identified on the plots in the last section are also identified on these plots. It is obvious that points having large influence on a particular parameter are points that have values at the extremes in partial regression plots. The values in parentheses correspond to the ranking according to the magnitude of $D_i(\hat{\beta}_j)$. The rankings for $\hat{\beta}_7$ and $\hat{\beta}_8$ are fairly informative. Although Jersey City is a very extreme point, the ranking of (5) and a value of

FIGURE 14. Partial regression plot associated with $\hat{\beta}_7$.

FIGURE 15. Partial regression plot associated with $\hat{\beta}_8$.

$|DFBETAS|$ below a liberal cutoff value indicate $\hat{\beta}_7$ is not affected much by presence or absence of Jersey City. However, in the plot of $\hat{\beta}_8$ Tampa falls in roughly the same location, but has a rank of (1), and had extremely large values for DFBETAS and $D_i(\hat{\beta}_8)$. The plot indicates that $\hat{\beta}_8$ would be much higher if Tampa was deleted. At least in these examples, the combination of $D_i(\hat{\beta}_j)$ (or DFBETAS) and partial regression plots is very informative. Although time-consuming, these diagnostics can be a great aid to the analyst in the study of his data and the fitted model.

4.8 INFLUENCE OF SUBSET OF β (See Section 2.9)

Sometimes the analyst might desire a diagnostic for a subset of the parameters being studied. In the model for the Lave and Seskin data, there are two logical subsets to study; the parameters associated with pollution variables, and the parameters associated with socioeconomic variables. Since the socioeconomic variables were more significant in the fitted model, this subset was chosen to use in the example for $D_i(\psi)$. The set $\psi = (\beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11})$. Using Equation 6, the full and reduced models were fit, and $D_i(\psi)$ was calculated for each observation. Appendix F contains a listing of the results. Figure 16 shows a stem-and-leaf display for the diagnostic.

The display shows two observations are definitely extreme points. These two observations are Tampa ($D_i = 1.61$) and Honolulu ($D_i = 1.14$). These two observations are more extreme when considered for only the socioeconomic variables than they were for all the variables (see Table 3). At this point it should be pointed out that the diagnostics are heavily dependent on what variables are in the regression model. Therefore, in the case of model building, diagnostics should always be done on the chosen model. Diagnostics can be used on preliminary models, but the reasonableness of the points (and model) should be done at the completion.

As an example of the dependency of diagnostics on the variables in the model, consider Figure 17. This figure is a plot of the hat matrix diagonal values for the reduced model ($\beta_0 - \beta_6$) versus the hat matrix diagonal values for the full model ($\beta_0 - \beta_{11}$). For the full model, there were four


```
STEM-AND-LEAF DISPLAY OF DI(SUB)
LEAF DIGIT UNIT = 0.0100
1 2 REPRESENTS 0.12
```

[illegible]

FIGURE 16. Stem-and-leaf display for $D_j(\psi)$.

observations with $h_i > 0.30$. The SMSAs corresponding to these four observations are identified on the plot. Three of the four observations have μ_i values (for the reduced model) of less than 0.10. If the study did not include the socioeconomic variables, Tampa and Honolulu would be much less important.

4.9 SUMMARY

In this chapter we have analyzed a reasonable-size data set, having 112 observations and 12 parameters (including β_0). Most of the diagnostics discussed in Chapter 2 were calculated and analyzed. The diagnostics showed that there were some highly influential values in the data. Given the results what should the next step be?

Most individuals would conclude that Tampa and Honolulu should be dropped from the model, the regression model fit to the 110 remaining

observations, and regression diagnostics calculated for the new regression fit. If these results seem reasonable, the analyst should try to determine why the two points are so extreme, and then report both sets of results.

The techniques discussed and used in this report are not intended as data-screening techniques, with observations that yield calculated values above the cutoff points being rejected. They are intended for the analyst to better understand the model and data under study.

FIGURE 17. Plot of hat matrix diagonals for reduced model ($\beta_1 - \beta_0$) vs. hat matrix diagonals for full model.

BIBLIOGRAPHY

- Abramowitz, M. and I. A. Stegun (eds.) 1964. National Bureau of Standards Applied Mathematics Series, 55: Handbook of Mathematical Functions, U.S. Government Printing Office, Washington, DC.
- Andrews, D. A. and D. Pregibon. 1978. Finding the outliers that matter. Journal of the Royal Statistical Society, Series B, 40:85-93.
- Atkinson, A. C. 1981. Two graphical displays for outlying and influential observations in regression. Biometrika 68:13-20.
- Belsley, D. A., E. Kuh and R. E. Welsch. 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley and Sons, New York.
- Blom, G. 1958. Statistical Estimates and Transformed Beta-Variables. John Wiley and Sons, New York.
- Cook, R. D. 1977. Detection of influential observations in linear regression. Technometrics, 19:15-18.
- Cook, R. D. 1979. Influential observations in linear regression. Journal of the American Statistical Association 74:169-174.
- Cook, R. D. and P. Prescott. 1981. On the accuracy of Bonferroni significance levels for detecting outliers in linear models. Technometrics 23:59-63.
- Cook, R. D. and S. Weisberg, 1980. Characterizations of an empirical influence function for detecting influential cases in regression. Technometrics 23:59-63.
- Daniel, C. 1959. Use of half-normal plots in interpreting factorial two-level experiments. Technometrics 1:311-341.
- Dongarra, J. J., C. B. Moler, J. R. Bunch and G. W. Stewart. 1979. Linpack User's Guide. SIAM, Philadelphia.
- Draper, N. R. and J. A. John. 1981. Influential observations and outliers in regression. Technometrics 23:21-26.
- Draper, N. R. and H. Smith. 1981. Applied Regression Analysis. Second Edition. John Wiley and Sons, New York.
- Ellenberg, J. H. 1976 Testing for a single outlier from a general linear regression. Biometrics 32:637-645.
- Filliben, J. J. 1975. The probability plot correlation coefficient test for normality. Technometrics 17:111-117.

- Hampel, F. R. 1974. The influence curve and its role in robust estimation. Journal of the American Statistical Association 69:383-393.
- Lave, L. B. and E. P. Seskin. 1977. Air Pollution and Human Health. The Johns Hopkins University Press, Baltimore.
- Lund, R. E. 1975. Tables for an approximate test for outliers in linear models. Technometrics 17:473-476.
- Prescott, P. 1975. An approximate test for outliers in linear models. Technometrics 17:129-132.
- Ryan, T. A. Jr. and B. L. Joiner. Normal Probability Plots and Tests for Normality. Technical Report, Statistics Department, Pennsylvania State Univ.
- Seber, G. A. F. 1977. Linear Regression Analysis. John Wiley and Sons, New York.
- Stewart, G. W. 1973. Introduction to Matrix Computations. Academic Press, New York.
- Tietjen, G. L. and R. H. Moore. 1972. Some Grubbs-Type statistics for the detection of several outliers. Technometrics 14:583-597.
- Tietjen, G. L., R. H. Moore and R. J. Beckman. 1973. Testing for a single outlier from a general linear regression. Biometrics 32:637-645.
- Velleman, P. F. and R. E. Welsch. 1981. Efficient computing of regression diagnostics in standard statistics packages. In 1980 Proceedings of the ASA Computing Section. American Statistical Association, Washington, DC.
- Weisberg, S. 1981. A statistic for allocating C_p to individual cases. Technometrics 14:583-597.

APPENDIX A

1969 POLLUTION-MORTALITY DATA

APPENDIX A

1969 POLLUTION-MORTALITY DATA

COLUMN COUNT ROW	ID 112	MIN-S 112	MEAN-S 112	MAX-S 112	MIN-P 112	MEAN-P 112	MAX-P 112
1	402.	33.	108.	274.	72.	134.	296.
2	403.	21.	83.	291.	16.	84.	150.
3	405.	11.	52.	133.	29.	123.	247.
4	407.	14.	65.	128.	16.	88.	178.
5	410.	9.	50.	137.	28.	106.	316.
6	417.	7.	51.	136.	13.	62.	160.
7	421.	9.	34.	76.	21.	88.	202.
8	422.	51.	124.	295.	11.	124.	603.
9	424.	67.	121.	204.	25.	69.	193.
10	426.	75.	200.	402.	30.	95.	207.
11	429.	99.	192.	420.	31.	127.	262.
12	431.	48.	136.	267.	1.	79.	621.
13	432.	22.	127.	219.	17.	69.	132.
14	438.	23.	68.	197.	24.	63.	129.
15	440.	44.	230.	741.	31.	118.	426.
16	445.	26.	100.	408.	24.	88.	207.
17	451.	33.	118.	248.	38.	118.	224.
18	452.	31.	94.	264.	25.	83.	174.
19	455.	9.	68.	209.	40.	103.	214.
20	457.	6.	46.	178.	13.	65.	105.
21	458.	13.	43.	92.	20.	67.	136.
22	460.	27.	69.	166.	30.	78.	179.
23	461.	32.	75.	148.	38.	75.	180.
24	462.	18.	63.	148.	28.	69.	150.
25	468.	61.	124.	213.	25.	58.	94.
26	475.	9.	85.	127.	31.	89.	352.
27	476.	13.	130.	464.	38.	99.	225.
28	479.	50.	126.	243.	42.	90.	259.
29	483.	30.	85.	233.	21.	73.	139.
30	484.	23.	64.	280.	22.	81.	262.
31	492.	19.	74.	267.	29.	115.	287.
32	493.	11.	43.	101.	28.	119.	236.
33	499.	46.	121.	253.	49.	89.	200.
34	500.	57.	112.	268.	35.	74.	166.
35	501.	30.	61.	97.	9.	76.	231.
36	510.	50.	191.	572.	19.	121.	354.
37	518.	37.	102.	184.	29.	100.	215.
38	519.	54.	102.	185.	62.	103.	190.
39	524.	39.	119.	294.	40.	99.	185.
40	527.	32.	144.	315.	37.	106.	222.
41	528.	35.	156.	348.	44.	101.	188.
42	532.	29.	102.	254.	27.	76.	135.
43	533.	29.	118.	253.	39.	132.	302.
44	536.	20.	72.	168.	18.	84.	200.
45	539.	32.	168.	284.	38.	103.	183.
46	542.	48.	102.	176.	21.	68.	101.
47	543.	23.	153.	261.	56.	120.	228.
48	548.	26.	99.	384.	45.	99.	233.
49	551.	35.	119.	214.	35.	80.	128.
50	553.	10.	50.	127.	23.	41.	59.
51	554.	87.	139.	182.	36.	81.	152.
52	556.	20.	139.	237.	60.	111.	202.
53	557.	53.	100.	173.	38.	106.	266.
54	558.	29.	64.	120.	28.	76.	154.
55	559.	38.	116.	263.	41.	98.	197.
56	565.	24.	56.	149.	30.	88.	192.

COLUMN COUNT ROW	ID 112	MIN-S 112	MEAN-S 112	MAX-S 112	MIN-P 112	MEAN-P 112	MAX-P 112
57	568.	19.	50.	88.	26.	53.	142.
58	571.	26.	70.	342.	32.	90.	299.
59	575.	42.	148.	552.	49.	94.	211.
60	577.	50.	103.	208.	51.	83.	134.
61	581.	26.	93.	161.	10.	75.	271.
62	582.	49.	258.	733.	56.	213.	493.
63	583.	44.	221.	874.	43.	182.	429.
64	585.	29.	95.	495.	27.	68.	268.
65	465.	68.	153.	268.	39.	92.	191.
66	544.	78.	220.	414.	37.	129.	386.
67	478.	37.	129.	294.	57.	129.	301.
68	443.	84.	188.	449.	16.	172.	456.
69	526.	54.	159.	325.	44.	128.	389.
70	485.	7.	67.	166.	30.	76.	161.
71	489.	74.	231.	496.	30.	155.	480.
72	525.	33.	151.	604.	12.	102.	385.
73	464.	20.	139.	368.	45.	118.	265.
74	436.	32.	84.	178.	22.	85.	166.
75	563.	16.	48.	110.	35.	84.	275.
76	430.	62.	139.	339.	38.	77.	120.
77	413.	24.	112.	252.	40.	104.	280.
78	420.	24.	67.	170.	25.	59.	137.
79	579.	29.	55.	151.	7.	75.	472.
80	419.	31.	83.	296.	37.	80.	204.
81	586.	34.	99.	241.	51.	124.	296.
82	590.	27.	77.	248.	18.	64.	133.
83	406.	19.	78.	164.	24.	85.	138.
84	418.	18.	131.	282.	39.	140.	255.
85	423.	53.	119.	214.	36.	72.	123.
86	434.	28.	98.	167.	26.	76.	118.
87	441.	28.	45.	71.	18.	42.	80.
88	788.	54.	155.	255.	53.	144.	353.
89	645.	18.	94.	265.	33.	81.	123.
90	447.	36.	124.	304.	31.	95.	210.
91	448.	45.	101.	263.	39.	87.	178.
92	449.	39.	142.	459.	30.	144.	504.
93	454.	22.	122.	444.	54.	170.	319.
94	459.	35.	122.	379.	51.	127.	244.
95	654.	60.	153.	474.	10.	89.	501.
96	661.	62.	127.	228.	40.	80.	149.
97	736.	47.	128.	416.	41.	77.	147.
98	502.	35.	94.	193.	15.	90.	286.
99	512.	52.	141.	553.	18.	104.	244.
100	514.	10.	163.	354.	21.	107.	320.
101	516.	16.	77.	213.	24.	86.	335.
102	776.	18.	64.	126.	54.	85.	204.
103	894.	30.	91.	183.	35.	106.	234.
104	692.	19.	56.	152.	18.	68.	133.
105	535.	20.	111.	425.	16.	94.	444.
106	541.	62.	156.	306.	10.	81.	211.
107	574.	61.	132.	220.	20.	67.	134.
108	580.	14.	36.	66.	19.	101.	409.
109	733.	21.	105.	477.	29.	78.	182.
110	587.	28.	87.	241.	10.	60.	127.
111	504.	59.	114.	193.	23.	131.	978.
112	567.	23.	91.	230.	47.	88.	128.

COLUMN COUNT ROW	ID 112	POP-DENS 112	G1E65 112	NON-WH 112	POOR 112	LOG-POP 112	MORT-RT 112
1	402.	25.50	77.	325.	186.	550.107	858.00
2	403.	21.30	86.	365.	192.	522.513	1086.00
3	405.	10.20	92.	52.	89.	597.180	801.00
4	407.	36.90	92.	207.	134.	545.143	943.00
5	410.	6.80	90.	79.	142.	561.102	865.00
6	417.	63.40	71.	98.	86.	579.143	768.00
7	421.	78.60	61.	53.	56.	600.965	574.00
8	422.	32.60	78.	52.	68.	607.835	745.00
9	424.	108.70	94.	69.	49.	590.527	858.00
10	426.	121.80	101.	78.	60.	586.716	940.00
11	429.	54.60	77.	133.	71.	564.167	891.00
12	431.	67.90	74.	230.	141.	571.727	904.00
13	432.	60.40	134.	153.	109.	609.150	1061.00
14	438.	21.40	57.	291.	187.	537.388	791.00
15	440.	42.30	80.	345.	169.	527.374	1096.00
16	445.	46.70	87.	65.	64.	538.549	814.00
17	451.	199.20	86.	170.	65.	589.366	930.00
18	452.	52.40	97.	77.	59.	538.810	925.00
19	455.	49.00	98.	44.	61.	545.351	915.00
20	457.	28.00	102.	81.	69.	518.729	877.00
21	458.	34.70	79.	85.	80.	554.400	771.00
22	460.	60.80	60.	291.	136.	544.662	736.00
23	461.	89.50	83.	320.	164.	598.715	984.00
24	462.	16.80	92.	328.	182.	546.743	976.00
25	468.	79.30	116.	14.	81.	564.318	1080.00
26	475.	50.30	106.	44.	67.	576.192	1007.00
27	476.	41.90	114.	12.	56.	580.106	1032.00
28	479.	68.10	66.	135.	69.	564.082	754.00
29	483.	26.60	78.	123.	77.	533.615	833.00
30	484.	3.60	121.	12.	83.	542.566	1258.00
31	492.	34.50	89.	73.	68.	572.584	841.00
32	493.	3.30	51.	103.	70.	541.273	678.00
33	499.	1302.40	113.	104.	91.	578.491	1213.00
34	500.	262.50	98.	188.	68.	626.478	984.00
35	501.	26.50	60.	42.	130.	549.194	633.00
36	510.	537.10	108.	171.	92.	706.010	1052.00
37	518.	65.20	65.	241.	99.	553.958	773.00
38	519.	43.40	76.	224.	103.	545.390	808.00
39	524.	64.00	98.	59.	58.	556.705	987.00
40	527.	152.10	77.	129.	76.	591.287	817.00
41	528.	62.70	75.	116.	60.	590.496	786.00
42	532.	140.40	104.	114.	66.	568.271	1072.00
43	533.	52.00	94.	97.	65.	572.697	968.00
44	536.	27.00	109.	36.	69.	599.581	992.00
45	539.	49.60	109.	13.	52.	573.112	1045.00
46	542.	35.20	101.	75.	66.	557.786	999.00
47	543.	14.90	112.	14.	97.	542.261	1161.00
48	548.	38.70	128.	6.	89.	553.500	1294.00
49	551.	124.20	117.	29.	80.	566.285	1052.00
50	553.	21.50	62.	267.	143.	550.061	734.00
51	554.	30.00	74.	167.	118.	537.561	876.00
52	556.	30.30	94.	161.	133.	548.109	976.00
53	557.	27.90	92.	72.	143.	559.891	890.00
54	558.	94.30	81.	371.	168.	565.279	909.00
55	559.	86.80	87.	198.	112.	564.658	897.00
56	565.	46.40	76.	111.	80.	587.110	790.00

COLUMN COUNT ROW	ID 112	POP-DENS 112	GIL65 112	NON-WH 112	POOR 112	LOG-POP 112	MORT-RT 112
57	568.	65.50	75.	74.	160.	591.176	745.00
58	571.	59.00	76.	19.	75.	565.422	675.00
59	575.	98.40	60.	259.	134.	582.633	769.00
60	577.	64.80	88.	261.	89.	567.542	998.00
61	581.	23.90	84.	66.	80.	560.432	862.00
62	582.	25.60	91.	57.	130.	536.522	1017.00
63	583.	18.10	108.	29.	148.	540.457	1044.00
64	585.	23.60	76.	18.	54.	545.248	691.00
65	465.	189.10	110.	44.	60.	652.490	1019.00
66	544.	134.20	97.	178.	73.	667.856	1016.00
67	478.	212.80	81.	182.	65.	661.869	866.00
68	443.	185.60	88.	182.	68.	683.905	977.00
69	526.	277.40	93.	174.	69.	628.019	976.00
70	485.	84.40	88.	26.	46.	625.054	778.00
71	489.	71.70	98.	165.	81.	635.853	1001.00
72	525.	155.80	102.	133.	81.	605.497	1050.00
73	464.	107.00	85.	248.	85.	628.611	965.00
74	436.	78.20	65.	225.	91.	613.125	789.00
75	563.	41.40	71.	162.	86.	616.827	767.00
76	430.	177.40	62.	264.	61.	642.190	717.00
77	413.	170.70	89.	120.	82.	691.811	858.00
78	420.	97.70	94.	164.	72.	650.918	901.00
79	579.	32.80	86.	58.	52.	614.314	861.00
80	419.	31.10	87.	73.	86.	612.234	759.00
81	586.	161.20	97.	86.	57.	610.596	909.00
82	590.	22.30	48.	155.	135.	525.435	582.00
83	406.	3.70	100.	60.	108.	553.538	841.00
84	418.	4.00	109.	57.	103.	604.525	901.00
85	423.	124.20	94.	73.	46.	589.148	871.00
86	434.	75.80	200.	111.	107.	599.501	1379.00
87	441.	103.40	49.	590.	72.	578.979	487.00
88	788.	24.40	97.	49.	57.	549.339	914.00
89	645.	18.20	117.	50.	67.	520.373	1099.00
90	447.	30.20	111.	59.	91.	531.017	957.00
91	448.	41.00	85.	70.	51.	544.033	812.00
92	449.	66.80	67.	178.	70.	579.753	822.00
93	454.	34.70	96.	39.	67.	548.499	928.00
94	459.	89.80	89.	123.	86.	591.191	957.00
95	654.	47.40	93.	58.	61.	560.880	865.00
96	661.	314.90	97.	58.	43.	612.762	900.00
97	736.	131.60	97.	166.	64.	547.743	967.00
98	502.	32.30	112.	35.	61.	585.459	1088.00
99	512.	103.60	98.	75.	52.	584.470	872.00
100	514.	26.00	94.	41.	71.	579.881	929.00
101	516.	12.70	113.	25.	74.	553.107	1077.00
102	776.	44.20	82.	328.	123.	511.598	884.00
103	694.	50.50	78.	226.	110.	532.604	886.00
104	692.	4.20	92.	5.	73.	507.486	712.00
105	535.	12.40	94.	112.	98.	567.316	899.00
106	541.	32.20	95.	35.	68.	541.869	1020.00
107	574.	112.30	50.	264.	101.	545.544	655.00
108	580.	16.30	111.	23.	86.	545.723	1045.00
109	733.	42.70	95.	18.	57.	506.515	816.00
110	587.	49.80	91.	64.	56.	522.513	837.00
111	504.	84.60	99.	85.	68.	612.872	1012.00
112	567.	98.20	59.	207.	98.	622.841	708.00

APPENDIX B

LISTING OF SMSA CODE NUMBERS AND
LOCATIONS FOR 1969 POLLUTION-MORTALITY DATA

APPENDIX B

LISTING OF SMSA CODE NUMBERS AND LOCATIONS FOR 1969 POLLUTION-MORTALITY DATA

402	Mobile, AL	461	New Orleans, LA	543	Johnstown, PA
403	Montgomery, AL	462	Shreveport, LA	544	Philadelphia, PA
405	Phoenix, AZ	464	Baltimore, MD	548	Wilkes-Barre, PA
406	Tucson, AZ	465	Boston, MA*	551	Providence, RI*
407	Little Rock, AR	468	Fall River, MA*	553	Columbia, SC
410	Fresno, CA	475	Springfield, MA*	554	Greenville, SC
413	Los Angeles, CA	476	Worcester, MA*	556	Chattanooga, TN
417	Sacramento, CA	478	Detroit, MI	557	Knoxville, TN
418	San Bernadino, CA	479	Flint, MI	558	Memphis, TN
419	San Diego, CA	483	Saginaw, MI	559	Nashville, TN
420	San Francisco, CA	484	Duluth, MN	563	Dallas, TX
421	San Jose, CA	485	Minneapolis, MN	565	Fort Worth, TX
422	Denver, CO	489	St. Louis, MO	567	Houston, TX
423	Bridgeport, CT*	492	Omaha, NB	568	San Antonio, TX
424	Hartford, CT*	493	Las Vegas, NV	571	Salt Lake City, UT
426	New Haven, CT*	499	Jersey City, NJ	574	Newport News, VA
429	Wilmington, DC	500	Newark, NJ	575	Norfolk, VA
430	Washington, DC	501	Albuquerque, NM	577	Richmond, VA
431	Jacksonville, FL	502	Albany, NY	579	Seattle, WA
423	Miami, FL	504	Buffalo, NY	580	Spokane, WA
434	Tampa, FL	510	New York, NY	581	Tacoma, WA
436	Atlanta, GA	512	Rochester, NY	582	Charleston, WV
438	Columbus, GA	514	Syracuse, NY	583	Huntington, WV
440	Savannah, GA	516	Utica, NY	585	Madison, WI
441	Honolulu, HI	518	Charlotte, NC	586	Milwaukee, WI
443	Chicago, IL	519	Greensboro, NC	587	Racine, WI
445	Rockford, IL	524	Canton, OH	590	Huntsville, AL
447	Evansville, IN	525	Cincinnati, OH	645	Springfield, IL
448	Fort Wayne, IN	526	Cleveland, OH	654	Grand Rapids, MI
449	Gary, IN	527	Columbus, OH	661	Paterson, NJ
451	Indianapolis, IN	528	Dayton, OH	692	Fargo, ND
452	South Bend, IN	532	Toledo, OH	733	Kenosha, WI
454	Davenport, IA	533	Youngstown, OH	736	Trenton, NJ
455	Des Moines, IA	535	Tulsa, OK	776	Durham, NC
457	Topeka, KS	538	Portland, OR	788	Peoria, IL
458	Wichita, KS	539	Allentown, PA	894	Winston-Salem, NC
459	Louisville, KY	541	Erie, PA		
460	Baton Rouge, LA	542	Harrisburg, PA		

*State Economic Area (SEA)

APPENDIX C

LISTING OF DIAGNOSTICS

APPENDIX C

LISTING OF DIAGNOSTICS

COLUMN COUNT	ID 112	H(1) 112	T(1) 112	R(1)* 112	D(1) 112	DIFFIT 112
ROW						
1	402.	0.174711	-1.34171	-1.34716	0.031745	-0.61984
2	403.	0.150505	2.10511	2.14257	0.065401	0.90184
3	405.	0.092625	-0.85188	-0.85070	0.006171	-0.27180
4	407.	0.071454	0.24703	0.24587	0.000391	0.06820
5	410.	0.079668	-0.48474	-0.48288	0.001694	-0.14207
6	417.	0.061006	0.73261	0.73090	0.002905	0.18630
7	421.	0.094154	-0.84598	-0.84477	0.006197	-0.27235
8	422.	0.104838	-1.30029	-1.30485	0.016495	-0.44655
9	424.	0.064771	-0.53682	-0.53490	0.001663	-0.14077
10	426.	0.082360	-0.82903	-0.82772	0.005138	-0.24797
11	429.	0.158300	0.51966	0.51776	0.004231	0.22454
12	431.	0.218513	0.07549	0.07512	0.000133	0.03972
13	432.	0.135805	-1.85196	-1.87512	0.044897	-0.74333
14	438.	0.103540	0.77437	0.77281	0.005769	0.26264
15	440.	0.237655	1.37853	1.38484	0.049349	0.77321
16	445.	0.051096	-0.53513	-0.53321	0.001284	-0.12373
17	451.	0.050163	0.58851	0.58658	0.001524	0.13480
18	452.	0.033325	0.18783	0.18692	0.000101	0.03471
19	455.	0.050447	0.07292	0.07256	0.000024	0.01672
20	457.	0.077160	-0.61063	-0.60870	0.002597	-0.17601
21	458.	0.038615	0.00996	0.00991	0.000000	0.00199
22	460.	0.047714	-0.20657	-0.20558	0.000178	-0.04602
23	461.	0.084180	1.23265	1.23590	0.011634	0.37470
24	462.	0.096151	0.07734	0.07696	0.000053	0.02310
25	468.	0.096861	0.72520	0.72347	0.004698	0.23693
26	475.	0.085639	0.42710	0.42535	0.001423	0.13017
27	476.	0.082371	0.28517	0.28385	0.000608	0.08504
28	479.	0.058186	-0.12096	-0.12037	0.000075	-0.02992
29	483.	0.034117	0.54080	0.53888	0.000861	0.10128
30	484.	0.077842	3.28836	3.46455	0.076035	1.00659
31	492.	0.045677	-0.20464	-0.20366	0.000167	-0.04456
32	493.	0.133541	1.09699	1.09812	0.015450	0.43111
33	499.	0.852658	-0.05020	-0.04995	0.001215	-0.12016
34	500.	0.066319	0.54096	0.53904	0.001731	0.14366
35	501.	0.133324	-0.84231	-0.84108	0.009092	-0.32988
36	510.	0.231627	-0.44787	-0.44607	0.005037	-0.24491
37	518.	0.061025	-0.02119	-0.02108	0.000002	-0.00537
38	519.	0.106934	-0.73656	-0.73487	0.005411	-0.25429
39	524.	0.028462	0.84548	0.84426	0.001744	0.14450
40	527.	0.047673	-0.29455	-0.29320	0.000362	-0.06560
41	528.	0.074737	-0.30099	-0.29961	0.000610	-0.08515
42	532.	0.030144	1.57787	1.58987	0.006446	0.28029
43	533.	0.046991	0.93403	0.93343	0.003583	0.20727
44	536.	0.044882	0.66212	0.66025	0.001716	0.14313
45	539.	0.110547	0.53302	0.53110	0.002941	0.18724
46	542.	0.049402	1.09215	1.09322	0.005164	0.24922
47	543.	0.125859	1.47411	1.48292	0.026062	0.56269
48	548.	0.096835	2.77955	2.87907	0.069002	0.94273
49	551.	0.049713	0.08556	0.08514	0.000032	0.01947
50	553.	0.079866	0.16439	0.16358	0.000195	0.04819
51	554.	0.153636	0.52228	0.52037	0.004125	0.22171
52	556.	0.118402	-0.26104	-0.25982	0.000762	-0.09522
53	557.	0.111321	-0.74967	-0.74802	0.005864	-0.26475
54	558.	0.094559	0.16594	0.16513	0.000240	0.05336
55	559.	0.020005	-0.32397	-0.32252	0.000178	-0.04608
56	565.	0.034577	0.43538	0.43361	0.000566	0.08206

COLUMN COUNT ROW	ID 112	H(1) 112	T(1) 112	R(1)* 112	D(1) 112	OFFIT 112
57	568.	0.169345	-0.55590	-0.55398	0.005248	-0.25013
58	571.	0.082259	-1.40604	-1.41303	0.014761	-0.42304
59	575.	0.132574	-0.06271	-0.06239	0.000050	-0.02439
60	577.	0.076990	1.35692	1.36272	0.012793	0.39357
61	581.	0.054345	0.40784	0.40614	0.000796	0.09736
62	582.	0.229419	-0.46925	-0.46742	0.005461	-0.25504
63	583.	0.253537	-1.10945	-1.11074	0.034825	-0.64734
64	585.	0.147497	-1.08932	-1.09035	0.017102	-0.45353
65	465.	0.094671	0.24590	0.24475	0.000527	0.07914
66	544.	0.110726	0.23599	0.23487	0.000578	0.08288
67	478.	0.093078	0.15377	0.15302	0.000202	0.04902
68	443.	0.272064	0.97467	0.97443	0.029576	0.59572
69	526.	0.056729	-0.04361	-0.04339	0.000010	-0.01064
70	485.	0.078936	-0.25733	-0.25613	0.000473	-0.07498
71	489.	0.110980	-0.42972	-0.42796	0.001920	-0.15121
72	525.	0.114393	0.90759	0.90678	0.008863	0.32590
73	464.	0.084433	1.04363	1.04410	0.008367	0.31707
74	436.	0.050504	1.03579	1.03618	0.004754	0.23897
75	563.	0.062516	0.42490	0.42315	0.001003	0.10927
76	430.	0.103505	-0.07084	-0.07049	0.000048	-0.02395
77	413.	0.118575	-0.12340	-0.12279	0.000171	-0.04504
78	420.	0.080915	0.67665	0.67481	0.003358	0.20022
79	579.	0.104625	0.78312	0.78159	0.005969	0.26718
80	419.	0.074871	-1.07477	-1.07561	0.007787	-0.30599
81	586.	0.058955	-0.33544	-0.33395	0.000587	-0.08359
82	590.	0.095835	-0.85908	-0.85795	0.006516	-0.27932
83	406.	0.040893	-1.38869	-1.39525	0.006849	-0.28810
84	418.	0.086870	-1.81299	-1.83430	0.026048	-0.56577
85	423.	0.049814	-0.27825	-0.27696	0.000338	-0.06341
86	434.	0.373175	-3.84122	-4.13954	0.731726	-3.19400
87	441.	0.361589	-3.40101	-3.59848	0.545728	-2.70817
88	788.	0.100688	-0.96328	-0.96293	0.008654	-0.32220
89	645.	0.060185	0.91497	0.91422	0.004466	0.23135
90	447.	0.034937	-1.27165	-1.27563	0.004876	-0.24271
91	448.	0.047973	-0.39138	-0.38972	0.000643	-0.08748
92	449.	0.086438	0.35676	0.35520	0.001003	0.10926
93	454.	0.133762	-0.20988	-0.20888	0.000567	-0.08208
94	459.	0.045180	0.85451	0.85335	0.002878	0.18563
95	654.	0.111846	-1.12477	-1.12628	0.013271	-0.39968
96	661.	0.074492	-0.43688	-0.43511	0.001280	-0.12344
97	736.	0.077283	0.17344	0.17260	0.000210	0.04995
98	502.	0.042712	1.46246	1.47095	0.007949	0.31071
99	512.	0.107190	-0.81407	-0.81269	0.006628	-0.28159
100	514.	0.165693	-0.00898	-0.00893	0.000001	-0.00398
101	516.	0.049552	0.97051	0.97023	0.004091	0.22153
102	776.	0.136086	-0.52536	-0.52345	0.003622	-0.20775
103	894.	0.047851	0.31259	0.31117	0.000409	0.06976
104	692.	0.060088	-2.19271	-2.23613	0.025604	-0.56539
105	535.	0.073237	-0.60208	-0.60015	0.002386	-0.16871
106	541.	0.093693	1.48402	1.49311	0.018965	0.48007
107	574.	0.104773	-0.60775	-0.60582	0.003601	-0.20725
108	580.	0.102776	0.70241	0.70062	0.004708	0.23713
109	733.	0.103785	-1.30231	-1.30691	0.016360	-0.44474
110	587.	0.067988	-0.35437	-0.35281	0.000763	-0.09529
111	504.	0.427724	-0.27048	-0.26923	0.004555	-0.23275
112	567.	0.081065	0.28335	0.28204	0.000590	0.08377

COLUMN COUNT ROW	ID 112	ATKIN-T 112	AN-PREG 112	COVRATIO 112	COOK-VOL 112
1	402.	1.78896	0.810432	1.09926	-0.04689
2	403.	2.60288	0.811849	0.77089	0.11695
3	405.	0.78446	0.900790	1.13929	-0.05933
4	407.	0.19685	0.927980	1.20613	-0.08450
5	410.	0.41005	0.918169	1.19172	-0.07936
6	417.	0.53769	0.933954	1.12634	-0.05267
7	421.	0.78606	0.899363	1.14259	-0.06073
8	422.	1.28882	0.880027	1.02709	-0.01238
9	424.	0.40628	0.932534	1.16525	-0.06840
10	426.	0.71570	0.911333	1.13178	-0.05584
11	429.	0.64805	0.839427	1.29756	-0.12209
12	431.	0.11464	0.781442	1.44264	-0.17376
13	432.	2.14538	0.834555	0.85879	0.06818
14	438.	0.75803	0.891084	1.17085	-0.07236
15	440.	2.23162	0.747858	1.17553	-0.08095
16	445.	0.35711	0.946187	1.14871	-0.06124
17	451.	0.38906	0.946547	1.13932	-0.05744
18	452.	0.10017	0.966334	1.16214	-0.06580
19	455.	0.04827	0.949503	1.18736	-0.07638
20	457.	0.50799	0.919399	1.16892	-0.07040
21	458.	0.00573	0.961384	1.17348	-0.07048
22	460.	0.13281	0.951880	1.17865	-0.07289
23	461.	1.08145	0.901904	1.02512	-0.01055
24	462.	0.07244	0.903795	1.24730	-0.10101
25	468.	0.68382	0.898390	1.17258	-0.07273
26	475.	0.37570	0.912693	1.20710	-0.08552
27	476.	0.24545	0.916883	1.21750	-0.08930
28	479.	0.08635	0.941676	1.19578	-0.07996
29	483.	0.29230	0.963059	1.12769	-0.05204
30	484.	2.90519	0.822442	0.30987	0.53810
31	492.	0.12860	0.953923	1.17625	-0.07187
32	493.	1.24425	0.856032	1.12599	-0.05587
33	499.	0.34679	0.147338	7.65457	-1.00815
34	500.	0.41463	0.930948	1.16656	-0.06898
35	501.	0.95210	0.860527	1.19513	-0.08318
36	510.	0.70686	0.766832	1.43332	-0.17149
37	518.	0.01551	0.938971	1.20143	-0.08225
38	519.	0.73392	0.888221	1.18343	-0.07742
39	524.	0.41706	0.964593	1.06544	-0.02577
40	527.	0.18933	0.931301	1.17238	-0.07044
41	528.	0.24576	0.924425	1.20612	-0.08465
42	532.	0.80897	0.945710	0.85956	0.07257
43	533.	0.59823	0.944694	1.06565	-0.02666
44	538.	0.41309	0.950931	1.12053	-0.04959
45	539.	0.54040	0.886926	1.22582	-0.09372
46	542.	0.71929	0.939260	1.02765	-0.01013
47	543.	1.62403	0.855146	0.99152	0.00278
48	548.	2.72087	0.833387	0.47596	0.34052
49	551.	0.05620	0.950218	1.18615	-0.07589
50	553.	0.13910	0.919885	1.22213	-0.09092
51	554.	0.63989	0.844055	1.28999	-0.11917
52	556.	0.27482	0.880998	1.26927	-0.11005
53	557.	0.76410	0.883684	1.18648	-0.07880
54	558.	0.15402	0.905192	1.24189	-0.09894
55	559.	0.13300	0.978966	1.13679	-0.05512
56	565.	0.23684	0.963593	1.14228	-0.05795

COLUMN COUNT ROW	ID 112	ATKIN-T 112	AN-PREG 112	COVRATIO 112	COOK-VOL 112
57	568.	0.72192	0.828088	1.30866	-0.12654
58	571.	1.22097	0.899598	0.96738	0.01611
59	575.	0.07040	0.867392	1.29999	-0.12169
60	577.	1.13591	0.906015	0.97796	0.01136
61	581.	0.28100	0.944082	1.16942	-0.06957
62	582.	0.73609	0.768884	1.42584	-0.16897
63	583.	1.86833	0.737275	1.30264	-0.12888
64	585.	1.30897	0.842387	1.14676	-0.06493
65	465.	0.22842	0.904782	1.23714	-0.09719
66	544.	0.23920	0.888778	1.26020	-0.10640
67	478.	0.14148	0.906708	1.24044	-0.09834
68	443.	1.71934	0.721021	1.38210	-0.15706
69	526.	0.03071	0.943253	1.19575	-0.07989
70	485.	0.21641	0.920455	1.21517	-0.08826
71	489.	0.43641	0.887378	1.24118	-0.09944
72	525.	0.94060	0.878313	1.15353	-0.06604
73	464.	0.91511	0.905595	1.08047	-0.03467
74	436.	0.68972	0.939309	1.04393	-0.01738
75	563.	0.31538	0.935791	1.17760	-0.07313
76	430.	0.06913	0.896450	1.25768	-0.10515
77	413.	0.12998	0.881291	1.27761	-0.11306
78	420.	0.57788	0.914877	1.16173	-0.06774
79	579.	0.77111	0.889884	1.17034	-0.07221
80	419.	0.88315	0.914443	1.06078	-0.02580
81	586.	0.24125	0.939986	1.18277	-0.07498
82	590.	0.80616	0.897492	1.14163	-0.06042
83	406.	0.83151	0.940610	0.93116	0.03543
84	418.	1.63290	0.883116	0.82731	0.08759
85	423.	0.18302	0.949451	1.17634	-0.07208
86	434.	9.21843	0.534337	0.26501	0.59368
87	441.	7.81625	0.564567	0.40426	0.40090
88	788.	0.92994	0.890967	1.12172	-0.05258
89	645.	0.66772	0.931947	1.08523	-0.03559
90	447.	0.70050	0.949457	0.96130	0.02109
91	448.	0.25249	0.950568	1.16343	-0.06694
92	449.	0.31534	0.912399	1.21619	-0.08899
93	454.	0.23690	0.865856	1.29552	-0.12017
94	459.	0.53575	0.947847	1.08209	-0.03360
95	654.	1.15355	0.876918	1.09028	-0.04007
96	661.	0.35627	0.923742	1.19135	-0.07899
97	736.	0.14417	0.922440	1.21826	-0.08935
98	502.	0.89675	0.936814	0.90920	0.04629
99	512.	0.81272	0.886893	1.16672	-0.07091
100	514.	0.01149	0.834306	1.35222	-0.14136
101	516.	0.63938	0.941496	1.05957	-0.02415
102	776.	0.59961	0.861530	1.26329	-0.10873
103	894.	0.20133	0.951219	1.17106	-0.06993
104	692.	1.63182	0.894721	0.66450	0.18923
105	535.	0.48692	0.923403	1.16542	-0.06885
106	541.	1.38558	0.886347	0.95290	0.02250
107	574.	0.59817	0.891921	1.20548	-0.08578
108	580.	0.68439	0.892797	1.18495	-0.07781
109	733.	1.28360	0.881016	1.02522	-0.01150
110	587.	0.27503	0.930841	1.19236	-0.07909
111	504.	0.67177	0.571857	1.95415	-0.32583
112	567.	0.24177	0.918198	1.21592	-0.08864

APPENDIX D

LISTING OF DFBETAS

APPENDIX D

LISTING OF DFBETAS

COLUMN COUNT ROW	C1 112	DFBETAS1 112	DFBETAS2 112	DFBETAS3 112	DFBETAS4 112	DFBETAS5 112	DFBETAS6 112
1	402.	-0.105280	0.136619	-0.041759	-0.360656	-0.003355	-0.125410
2	403.	0.062373	-0.165150	0.156756	-0.417663	0.286745	-0.257196
3	405.	-0.018519	0.101300	0.014421	0.086903	-0.193135	0.107333
4	407.	-0.010345	0.002271	-0.019862	-0.043744	0.034809	-0.022878
5	410.	-0.002802	0.035371	0.011488	0.009658	-0.035250	-0.009591
6	417.	-0.085808	0.038312	-0.024755	-0.064428	-0.015872	-0.025324
7	421.	0.048938	0.029753	0.035426	0.068335	-0.084277	0.060702
8	422.	-0.075078	0.029168	0.049355	0.115318	-0.033210	-0.211996
9	424.	-0.078198	0.006116	0.013390	-0.008102	0.035430	0.001025
10	426.	-0.037443	-0.116264	0.055682	0.020939	0.049199	0.041008
11	429.	0.159828	-0.041609	0.009065	-0.055588	0.086173	-0.077984
12	431.	-0.006380	0.014326	-0.010725	-0.004525	-0.017559	0.027663
13	432.	0.291091	-0.342086	0.219077	0.208980	0.103274	0.114244
14	438.	0.008680	-0.005248	0.019588	-0.022082	-0.032515	-0.026927
15	440.	-0.248822	0.275303	0.147364	0.003555	-0.243717	0.214766
16	445.	-0.002621	0.040161	-0.077468	0.016256	-0.001341	0.017735
17	451.	-0.033071	0.018262	-0.042000	-0.032778	0.074973	-0.040479
18	452.	0.002087	-0.004184	0.003733	-0.007762	0.004821	-0.007012
19	455.	-0.004506	-0.002628	-0.000018	0.003808	0.003238	0.000298
20	457.	0.032178	0.022277	-0.012899	0.080366	-0.042407	0.048904
21	458.	-0.000371	-0.000043	-0.000393	-0.000541	0.000214	-0.000444
22	460.	-0.001448	0.004388	0.002363	0.002327	-0.002769	0.000766
23	461.	0.072274	-0.078254	0.023272	0.095784	-0.065564	0.028878
24	462.	0.000514	-0.002592	0.000014	-0.001234	-0.000774	-0.000242
25	468.	0.105348	0.012289	-0.019733	0.008982	-0.067481	-0.040685
26	475.	-0.081436	0.064395	-0.067322	0.033993	-0.050510	0.075697
27	476.	-0.040072	0.009956	0.032290	0.025048	-0.028861	0.005879
28	479.	-0.000211	-0.009469	0.007967	-0.015376	0.013239	-0.010581
29	483.	0.001047	0.000761	0.003311	-0.035251	0.009181	-0.030985
30	484.	0.364759	-0.614461	0.505531	0.034721	-0.036331	0.126356
31	492.	-0.003882	0.024380	-0.013166	0.009860	-0.026104	0.008123
32	493.	-0.036031	-0.055700	-0.092036	-0.137910	0.271880	-0.118968
33	499.	0.006473	-0.000715	0.009030	0.000058	-0.002373	-0.007250
34	500.	0.069784	-0.057553	0.048995	0.029642	-0.026193	-0.007824
35	501.	-0.041327	-0.014536	0.077218	0.119488	-0.039783	0.047973
36	510.	0.027437	0.001974	-0.060558	0.087842	-0.028010	0.039490
37	518.	0.000101	-0.000650	0.002152	0.002380	-0.003186	0.001724
38	519.	-0.117731	0.067765	0.004031	-0.165498	0.014763	-0.033721
39	524.	0.017849	-0.002404	0.003161	0.043845	0.002006	-0.023055
40	527.	0.041811	-0.045870	0.026350	0.000930	0.004337	0.007362
41	528.	0.046030	-0.055798	0.022103	-0.024891	0.027116	0.002124
42	532.	-0.063464	0.038125	-0.002279	-0.052366	-0.014303	-0.070338
43	533.	-0.045566	0.025975	-0.079245	-0.027197	0.121810	-0.026652
44	538.	-0.004796	-0.022058	-0.005954	-0.055746	0.034645	-0.042091
45	539.	-0.110388	0.156003	-0.111487	0.007126	-0.032504	-0.013146
46	542.	0.092910	0.008606	-0.049347	-0.073227	0.016950	-0.100529
47	543.	-0.270872	0.363155	-0.304357	0.223821	-0.112304	0.068641
48	548.	0.249797	-0.462448	0.519324	0.451918	-0.178853	0.115788
49	551.	-0.002293	0.006816	-0.005842	0.003171	-0.005188	-0.003483
50	553.	-0.015791	0.010300	-0.001834	0.001182	-0.018996	-0.001717
51	554.	0.148443	-0.007366	-0.043386	0.019773	-0.006914	-0.027076
52	556.	0.048540	-0.058257	0.046024	-0.048512	0.027286	-0.017228
53	557.	-0.156305	0.072510	0.023048	-0.036047	-0.032952	0.012257
54	558.	0.009891	-0.012996	0.000075	-0.008755	0.009602	-0.008065
55	559.	-0.004094	-0.002845	0.005361	-0.014306	-0.000205	0.003941
56	565.	0.019075	-0.040761	0.011503	-0.000247	0.023295	-0.019391

COLUMN COUNT ROW	C1 112	DFBETAS1 112	DFBETAS2 112	DFBETAS3 112	DFBETAS4 112	DFBETAS5 112	DFBETAS6 112
57	568.	-0.007976	-0.003938	0.008604	-0.040021	0.082930	-0.004007
58	571.	-0.116027	0.245371	-0.271240	-0.144435	0.090603	-0.093600
59	575.	-0.001279	0.003748	-0.014954	-0.011341	0.010244	-0.002029
60	577.	0.149552	-0.095639	0.035960	0.213422	-0.064441	0.023059
61	581.	-0.037915	0.052997	-0.052196	-0.044317	-0.012366	0.011691
62	582.	0.041937	-0.065458	0.012153	0.017177	-0.095622	0.030034
63	583.	-0.097446	0.167448	-0.365879	0.047754	-0.168663	0.118051
64	585.	-0.044201	0.149693	-0.322929	-0.154288	0.227366	-0.124198
65	465.	0.030249	0.000556	-0.004298	0.017389	-0.014521	-0.011818
66	544.	0.004087	0.032456	-0.021863	0.000781	-0.007194	0.004743
67	478.	-0.006356	-0.002428	0.000622	0.022475	0.001154	0.004062
68	443.	0.252338	-0.176774	0.047355	-0.348649	0.420220	-0.230777
69	526.	0.000545	-0.001735	0.002680	-0.002107	-0.000516	-0.003289
70	485.	0.035866	-0.012136	0.003951	-0.010780	0.016423	0.004191
71	489.	-0.003698	-0.052580	0.037177	0.045003	-0.036523	0.000133
72	525.	-0.039908	-0.057862	0.207548	-0.099463	-0.032425	0.029504
73	464.	-0.190592	0.096787	-0.005316	0.054473	-0.013613	0.010061
74	436.	-0.004469	-0.010873	-0.017611	-0.103849	0.079054	-0.104241
75	563.	-0.012494	-0.023080	0.002223	0.042550	-0.023334	0.034870
76	430.	-0.004803	0.001504	-0.006213	-0.004176	0.004937	0.004750
77	413.	0.012720	-0.004400	0.000147	-0.010664	0.008285	-0.002116
78	420.	-0.007095	-0.036932	0.046349	0.018770	-0.053923	-0.015689
79	579.	0.021080	-0.061771	0.017878	-0.010509	-0.072539	0.157483
80	419.	-0.082429	0.136367	-0.168931	-0.144027	0.110931	-0.015896
81	586.	-0.014169	0.033147	-0.006943	-0.037932	-0.019191	-0.007586
82	590.	-0.001895	-0.013807	-0.027554	0.053812	0.030736	0.051444
83	406.	0.023422	-0.024628	0.074484	0.102719	-0.080987	0.131384
84	418.	0.190992	-0.139967	0.177989	0.117766	-0.275600	0.216991
85	423.	-0.015599	-0.008195	0.008353	-0.019517	0.021640	0.006703
86	434.	-0.385067	0.420446	0.013405	0.282633	-0.221529	0.337945
87	441.	0.270891	0.081823	0.053867	0.456984	-0.152896	0.032259
88	788.	-0.031259	-0.068620	0.155535	-0.086086	-0.074486	-0.041788
89	645.	-0.050156	0.019401	0.002998	0.016847	-0.014118	-0.024508
90	447.	-0.010455	-0.040615	0.028283	0.010448	0.002171	0.019080
91	448.	-0.034905	0.023386	-0.016054	-0.033288	0.009631	0.000497
92	449.	-0.009002	-0.014281	0.021246	-0.016214	0.031702	0.030298
93	454.	-0.004920	0.031442	-0.021350	-0.003195	-0.052886	0.020789
94	459.	0.039158	-0.085034	0.083950	0.076408	0.051717	-0.037700
95	654.	-0.081466	0.016581	-0.113332	0.015621	0.172717	-0.229341
96	661.	-0.043710	0.005574	0.007076	-0.032718	0.023460	0.011440
97	736.	0.011018	-0.011984	0.025794	0.019415	-0.017347	-0.003280
98	502.	0.046322	-0.026538	-0.058783	-0.142272	0.067945	-0.010383
99	512.	-0.117979	0.150750	-0.210693	0.098158	-0.071442	0.097685
100	514.	0.003387	-0.003356	0.001863	0.000812	0.000726	-0.000393
101	516.	-0.049428	-0.001727	-0.012234	0.031404	-0.064965	0.119753
102	776.	0.024164	0.004183	0.026089	-0.101935	0.035400	-0.076047
103	894.	-0.000791	0.000779	-0.023734	-0.011271	0.031780	-0.005005
104	692.	-0.016588	0.032453	0.050919	0.151494	-0.059383	0.110506
105	535.	0.047581	0.005327	-0.064630	0.003990	0.066096	-0.098216
106	541.	0.073202	0.184401	-0.144444	-0.237565	0.007375	-0.105511
107	574.	-0.015169	-0.078935	0.067091	0.060823	0.011291	0.036321
108	580.	0.026344	-0.072546	-0.044913	-0.032632	0.058563	0.080761
109	733.	0.035670	0.079909	-0.257948	-0.068409	0.123356	-0.024066
110	587.	0.011335	-0.014383	0.002858	0.045145	0.000425	0.020704
111	504.	-0.028386	0.015628	0.035493	-0.059225	0.065576	-0.204972
112	567.	-0.019411	0.004578	0.007260	0.033944	-0.016833	-0.010951

COLUMN COUNT	C1 112	DEBETAS7 112	DEBETAS8 112	DEBETAS9 112	DEBETAS10 112	DEBETAS11 112
ROW						
1	402.	0.079240	-0.05386	-0.06527	-0.23082	0.013455
2	403.	0.043332	0.25805	0.35929	0.29825	-0.165777
3	405.	0.015118	-0.00569	0.04328	-0.03674	-0.076209
4	407.	0.009038	0.01522	0.01863	0.01250	-0.014976
5	410.	-0.004529	0.00443	0.05779	-0.08989	-0.019871
6	417.	0.023817	-0.09209	-0.05130	0.01298	0.039134
7	421.	-0.035130	0.14843	0.07609	0.03159	-0.070303
8	422.	0.022214	0.09179	0.06867	-0.00293	-0.039286
9	424.	0.018294	-0.00385	0.02498	0.01766	-0.006619
10	426.	0.023094	0.00580	0.03295	0.01678	0.014768
11	429.	-0.027321	-0.01691	-0.00339	-0.00728	-0.040506
12	431.	0.002462	-0.00449	0.00122	0.00849	-0.004862
13	432.	0.113415	-0.36801	-0.13488	-0.09182	-0.179256
14	438.	0.005211	-0.08430	-0.04356	0.17789	0.013997
15	440.	-0.014835	-0.00857	0.22483	0.08299	-0.190155
16	445.	-0.004186	0.00937	-0.00141	0.03973	0.034991
17	451.	0.050803	-0.00348	0.05765	-0.06924	-0.028475
18	452.	0.000974	0.00764	0.00780	-0.01898	-0.018834
19	455.	0.001040	0.00204	0.00113	-0.00717	-0.005442
20	457.	-0.021172	-0.04981	-0.04756	0.06649	0.078888
21	458.	0.000153	-0.00059	-0.00040	-0.00015	-0.000169
22	460.	-0.004619	0.01266	-0.01345	-0.00724	0.010798
23	461.	-0.046663	0.06778	0.05354	0.19999	0.142097
24	462.	-0.002185	0.00912	0.00782	0.01161	-0.000299
25	468.	-0.031648	0.05357	-0.09534	0.06931	0.006529
26	475.	-0.001268	0.01159	-0.00166	-0.02827	-0.013236
27	476.	-0.015070	0.01382	-0.00519	-0.01942	0.012821
28	479.	0.002770	0.01537	0.00402	0.00686	0.006092
29	483.	0.001435	-0.02298	0.00918	-0.03246	-0.047522
30	484.	-0.135078	0.58167	-0.06204	0.06142	-0.102239
31	492.	0.000684	-0.00267	-0.00422	0.01255	0.000545
32	493.	0.057045	-0.19363	-0.00298	-0.10688	-0.106480
33	499.	-0.116196	0.00176	0.01127	-0.00967	0.035237
34	500.	0.019140	0.04279	0.03019	-0.01776	0.047047
35	501.	-0.044534	0.17987	0.21623	-0.19543	-0.017053
36	510.	-0.117395	-0.01282	0.00014	-0.04770	-0.113841
37	518.	-0.000681	0.00126	-0.00217	0.00136	0.002053
38	519.	0.046541	-0.00940	-0.03461	0.01484	0.056468
39	524.	-0.017461	0.01594	-0.00511	-0.06184	-0.046286
40	527.	-0.014392	0.03517	0.00674	0.01069	-0.001644
41	528.	0.012901	0.04148	0.00805	0.02412	-0.007699
42	532.	0.057415	0.10081	0.08805	-0.12603	-0.065021
43	533.	-0.001372	0.01679	0.05481	-0.09401	-0.054854
44	538.	-0.028397	0.04579	-0.01839	0.00053	0.056385
45	539.	-0.016502	-0.01358	-0.02396	-0.04134	-0.023297
46	542.	-0.044437	0.05386	-0.01201	-0.03312	-0.042819
47	543.	-0.049939	-0.00562	-0.15523	0.06096	-0.114189
48	548.	-0.166978	0.49803	-0.15489	0.14228	0.034535
49	551.	-0.000334	0.00476	-0.00774	0.00464	0.004038
50	553.	-0.002041	-0.01603	-0.00021	0.01443	0.004080
51	554.	-0.031755	-0.02662	-0.05116	0.06811	-0.040681
52	556.	0.008507	0.00772	0.01007	-0.01985	0.012106
53	557.	0.024696	0.00267	0.14306	-0.19095	-0.026228
54	558.	-0.000275	0.01426	0.02145	0.01986	0.009323
55	559.	0.001740	-0.00432	-0.01013	-0.00764	0.007303
56	565.	-0.008506	-0.02196	-0.01515	-0.00005	0.026894

COLUMN COUNT ROW	C1 112	DFBETAS7 112	DFBETAS8 112	DFBETAS9 112	DFBETAS10 112	DFBETAS11 112
57	566.	0.001816	0.08812	0.17916	-0.20330	-0.115359
58	571.	0.003498	0.12647	0.18956	-0.05322	-0.062648
59	575.	0.002625	0.00787	0.00227	-0.00672	-0.006494
60	577.	-0.103016	0.12079	0.18573	-0.11358	-0.047219
61	581.	0.004364	-0.02937	-0.02123	-0.00126	-0.016508
62	582.	-0.009002	0.04515	0.04503	-0.05486	0.058078
63	583.	0.020244	-0.03770	0.15722	-0.26680	-0.003218
64	585.	0.045841	0.10645	0.10805	0.05171	-0.001997
65	465.	-0.010244	0.00787	-0.02937	0.01697	0.045018
66	544.	-0.019625	0.00224	0.00694	-0.00033	0.036097
67	478.	0.000570	-0.00695	0.00515	-0.00904	0.025361
68	443.	0.009614	0.04858	0.10001	-0.00878	0.177885
69	526.	-0.003682	-0.00063	-0.00291	0.00281	-0.000677
70	485.	0.005845	0.02138	0.02190	0.01163	-0.039880
71	489.	0.030415	-0.01220	-0.02960	0.00815	-0.025660
72	525.	0.031379	0.05688	0.06355	-0.02402	0.058820
73	464.	-0.036073	-0.00515	0.14360	-0.10056	0.102752
74	436.	-0.015543	-0.08328	0.04336	-0.01359	0.101688
75	563.	-0.022159	-0.02752	-0.00505	-0.00134	0.056153
76	430.	0.002967	0.00678	-0.00456	0.00576	-0.010342
77	413.	0.003957	0.00812	0.01077	-0.00936	-0.038800
78	420.	-0.049287	0.02598	0.01493	-0.00636	0.151694
79	579.	-0.030788	-0.00343	-0.01115	-0.04078	0.063776
80	419.	0.102916	0.03411	0.12750	-0.09011	-0.198990
81	586.	-0.005181	-0.01226	-0.00398	0.02227	-0.015683
82	590.	-0.028615	0.18953	0.12295	-0.12582	0.022881
83	406.	0.023405	-0.02297	0.10961	-0.12255	-0.007060
84	418.	0.099152	-0.05699	0.09919	-0.13857	-0.180261
85	423.	0.006810	0.00297	0.01048	0.01716	-0.004493
86	434.	0.483714	-2.99454	-0.98054	-0.16626	-0.267895
87	441.	0.043365	-0.35033	-2.38467	1.68820	0.516354
88	788.	0.035330	-0.01122	-0.00583	0.09620	0.137377
89	645.	-0.016681	0.10965	0.04503	-0.08819	-0.116713
90	447.	0.017346	-0.10046	0.01796	-0.00919	0.126850
91	448.	0.013455	0.00671	0.00390	0.03723	0.030470
92	449.	0.005513	-0.02351	0.04021	-0.04628	-0.021885
93	454.	-0.001286	-0.00657	-0.00647	0.01950	0.017113
94	459.	-0.022265	0.00155	-0.00529	-0.00115	0.046627
95	654.	0.028274	-0.01211	0.00561	0.05608	0.082734
96	661.	-0.042339	0.00745	0.03355	0.01189	-0.014921
97	736.	-0.001014	0.01344	0.01734	-0.02098	-0.014776
98	502.	-0.049435	0.14133	0.00777	-0.05268	0.010597
99	512.	0.008651	-0.04682	-0.02870	0.05588	-0.020998
100	514.	-0.000034	0.00104	0.00036	0.00050	0.000165
101	516.	-0.020677	0.09162	-0.00309	-0.03320	-0.049394
102	776.	0.002758	-0.05081	-0.11801	0.05856	0.101327
103	894.	0.008517	0.00331	0.03342	-0.01485	-0.042133
104	692.	-0.056378	0.03681	0.15501	0.04957	0.292330
105	535.	0.015026	-0.01493	-0.01360	-0.00172	0.000819
106	541.	-0.014643	-0.04387	-0.09684	-0.00626	-0.156418
107	574.	-0.027481	0.09942	-0.02034	0.01329	0.066896
108	580.	0.016936	0.08940	0.00283	-0.00339	-0.063516
109	733.	-0.018316	0.00332	0.02232	0.12957	0.174461
110	587.	-0.012106	0.00229	-0.00809	0.03777	0.048649
111	504.	0.003921	-0.03285	-0.01767	0.01812	0.019460
112	567.	-0.010328	-0.04446	-0.01380	0.00978	0.048142

APPENDIX E

LISTING OF COOK'S $D_i(\hat{\beta}_j)$

APPENDIX E

LISTING OF COOK'S $D_i(\hat{\beta}_j)$

COLUMN COUNT ROW	C1 112	COOK-B1 112	COOK-B2 112	COOK-B3 112	COOK-B4 112	COOK-B5 112	COOK-B6 112
1	402.	0.010994	0.018514	0.001730	0.129022	0.000011	0.0156006
2	403.	0.003756	0.026329	0.023721	0.168396	0.079373	0.0638568
3	405.	0.000344	0.010290	0.000209	0.007573	0.037405	0.0115524
4	407.	0.000108	0.000005	0.000398	0.001932	0.001223	0.0005284
5	410.	0.000008	0.001261	0.000133	0.000094	0.001252	0.0000927
6	417.	0.007398	0.001475	0.000616	0.004170	0.000253	0.0006443
7	421.	0.002402	0.000888	0.001259	0.004683	0.007123	0.0036953
8	422.	0.005597	0.000845	0.002419	0.013205	0.001095	0.0446289
9	424.	0.006159	0.000038	0.000181	0.000066	0.001264	0.0000011
10	426.	0.001406	0.013560	0.003110	0.000440	0.002428	0.0016870
11	429.	0.025733	0.001744	0.000083	0.003113	0.007480	0.0061264
12	431.	0.000041	0.000207	0.000116	0.000021	0.000311	0.0007729
13	432.	0.082654	0.114151	0.046817	0.042601	0.010404	0.0127314
14	438.	0.000076	0.000028	0.000385	0.000490	0.001061	0.0007280
15	440.	0.061349	0.075102	0.021519	0.000013	0.058858	0.0457050
16	445.	0.000007	0.001625	0.006045	0.000266	0.000002	0.0003168
17	451.	0.001101	0.000336	0.001776	0.001082	0.005658	0.0016493
18	452.	0.000004	0.000018	0.000014	0.000061	0.000023	0.0000497
19	455.	0.000021	0.000007	0.000000	0.000015	0.000011	0.0000001
20	457.	0.001042	0.000499	0.000167	0.006500	0.001810	0.0024067
21	458.	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000002
22	460.	0.000002	0.000019	0.000006	0.000005	0.000008	0.0000006
23	461.	0.005196	0.006092	0.000539	0.009126	0.004276	0.0008295
24	462.	0.000000	0.000007	0.000000	0.000002	0.000001	0.0000001
25	468.	0.011151	0.000152	0.000391	0.000081	0.004575	0.0016632
26	475.	0.006687	0.004181	0.004570	0.001165	0.002572	0.0057774
27	476.	0.001621	0.000100	0.001052	0.000633	0.000841	0.0000349
28	479.	0.000000	0.000091	0.000064	0.000239	0.000177	0.0001131
29	483.	0.000001	0.000001	0.000011	0.001252	0.000085	0.0009670
30	484.	0.119861	0.340137	0.230230	0.001086	0.001189	0.0143831
31	492.	0.000015	0.000600	0.000175	0.000098	0.000688	0.0000666
32	493.	0.001296	0.003096	0.008453	0.018980	0.073767	0.0141242
33	499.	0.000042	0.000001	0.000082	0.000000	0.000006	0.0000531
34	500.	0.004905	0.003336	0.002418	0.000885	0.000691	0.0000616
35	501.	0.001713	0.000212	0.005980	0.014319	0.001587	0.0023082
36	510.	0.000759	0.000004	0.003697	0.007779	0.000791	0.0015720
37	518.	0.000000	0.000000	0.000005	0.000006	0.000010	0.0000030
38	519.	0.013925	0.004613	0.000016	0.027516	0.000219	0.0011423
39	524.	0.000319	0.000006	0.000010	0.001928	0.000004	0.0005331
40	527.	0.001764	0.002123	0.000701	0.000001	0.000019	0.0000547
41	528.	0.002138	0.003142	0.000493	0.000625	0.000742	0.0000046
42	532.	0.003967	0.001432	0.000005	0.002701	0.000201	0.0048730
43	533.	0.002079	0.000676	0.006288	0.000741	0.014857	0.0007112
44	538.	0.000023	0.000489	0.000036	0.003125	0.001207	0.0017817
45	539.	0.012274	0.024513	0.012519	0.000051	0.001064	0.0001741
46	542.	0.008615	0.000074	0.002430	0.005352	0.000287	0.0100865
47	543.	0.072502	0.130319	0.091536	0.049502	0.012463	0.0046558
48	548.	0.058160	0.199329	0.251375	0.190355	0.029815	0.0124959
49	551.	0.000005	0.000047	0.000034	0.000010	0.000027	0.0000123
50	553.	0.000252	0.000107	0.000003	0.000001	0.000364	0.0000030
51	554.	0.022197	0.000055	0.001896	0.000394	0.000048	0.0007385
52	556.	0.002378	0.003426	0.002138	0.002376	0.000752	0.0002996
53	557.	0.024539	0.005281	0.000534	0.001305	0.001091	0.0001509
54	558.	0.000099	0.000171	0.000000	0.000077	0.000093	0.0000657
55	559.	0.000017	0.000008	0.000029	0.000207	0.000000	0.0000157
56	565.	0.000367	0.001675	0.000133	0.000000	0.000547	0.0003791

COLUMN COUNT ROW	C1 112	COOK-B1 112	COOK-B2 112	COOK-B3 112	COOK-B4 112	COOK-B5 112	COOK-B6 112
57	568.	0.000064	0.000016	0.000075	0.001613	0.006925	0.0000162
58	571.	0.013329	0.059613	0.072845	0.020656	0.008128	0.0087116
59	575.	0.000002	0.000014	0.000226	0.000130	0.000106	0.0000042
60	577.	0.022176	0.009069	0.001282	0.045162	0.004117	0.0005272
61	581.	0.001450	0.002832	0.002747	0.001980	0.000154	0.0001426
62	582.	0.001773	0.004318	0.000149	0.000297	0.009216	0.0009091
63	583.	0.009474	0.027973	0.133555	0.002275	0.028381	0.0139035
64	585.	0.001950	0.022366	0.104087	0.023760	0.051598	0.0153961
65	465.	0.000924	0.000000	0.000019	0.000305	0.000213	0.0001410
66	544.	0.000017	0.001063	0.000483	0.000001	0.000052	0.0000227
67	478.	0.000041	0.000006	0.000000	0.000510	0.000001	0.0000167
68	443.	0.063707	0.031265	0.002244	0.121618	0.176674	0.0532851
69	526.	0.000000	0.000003	0.000007	0.000004	0.000000	0.0000109
70	485.	0.001298	0.000149	0.000016	0.000117	0.000272	0.0000177
71	489.	0.000014	0.002787	0.001394	0.002042	0.001345	0.0000000
72	525.	0.001595	0.003354	0.043153	0.009910	0.001053	0.0008721
73	464.	0.036293	0.009359	0.000028	0.002965	0.000185	0.0001011
74	436.	0.000020	0.000118	0.000310	0.010777	0.006245	0.0108581
75	563.	0.000157	0.000537	0.000005	0.001825	0.000549	0.0012260
76	430.	0.000023	0.000002	0.000039	0.000018	0.000025	0.0000228
77	413.	0.000163	0.000020	0.000000	0.000115	0.000069	0.0000045
78	420.	0.000051	0.001371	0.002160	0.000354	0.002924	0.0002475
79	579.	0.000446	0.003831	0.000321	0.000111	0.005283	0.0248977
80	419.	0.006784	0.018567	0.028493	0.020711	0.012286	0.0002523
81	586.	0.000203	0.001109	0.000049	0.001452	0.000372	0.0000561
82	590.	0.000004	0.000191	0.000761	0.002903	0.000947	0.0026535
83	406.	0.000543	0.000601	0.005496	0.010452	0.006497	0.0170998
84	418.	0.035635	0.019138	0.030948	0.013548	0.074201	0.0459973
85	423.	0.000246	0.000068	0.000070	0.000384	0.000473	0.0000453
86	434.	0.127675	0.152214	0.030155	0.068783	0.042257	0.0983390
87	441.	0.065549	0.005980	0.002592	0.186544	0.020882	0.0009295
88	788.	0.000978	0.004712	0.024209	0.007416	0.005552	0.0017475
89	645.	0.002520	0.000377	0.000009	0.000284	0.000200	0.0006016
90	447.	0.000109	0.001639	0.000795	0.000108	0.000005	0.0003618
91	448.	0.001229	0.000552	0.000260	0.001118	0.000094	0.0000002
92	449.	0.000082	0.000206	0.000455	0.000265	0.001014	0.0009261
93	454.	0.000024	0.000998	0.000460	0.000010	0.002824	0.0004364
94	459.	0.001538	0.007250	0.007067	0.005854	0.002682	0.0014251
95	654.	0.006619	0.000274	0.012810	0.000243	0.029751	0.0524563
96	661.	0.001926	0.000031	0.000050	0.001079	0.000555	0.0001319
97	736.	0.000123	0.000145	0.000672	0.000381	0.000304	0.0000109
98	502.	0.002308	0.000696	0.003416	0.020009	0.004563	0.0001066
99	512.	0.013966	0.022803	0.044543	0.009668	0.005121	0.0095749
100	514.	0.000012	0.000011	0.000004	0.000001	0.000001	0.0000002
101	516.	0.002445	0.000003	0.000150	0.000987	0.004223	0.0143492
102	776.	0.000588	0.000018	0.000686	0.010467	0.001262	0.0058254
103	894.	0.000001	0.000001	0.000568	0.000128	0.001019	0.0000253
104	692.	0.000265	0.001013	0.002493	0.022068	0.003391	0.0117419
105	535.	0.002279	0.000029	0.004204	0.000016	0.004397	0.0097084
106	541.	0.005293	0.033591	0.020611	0.055752	0.000054	0.0109974
107	574.	0.000232	0.006270	0.004530	0.003723	0.000128	0.0013276
108	580.	0.000698	0.005290	0.002028	0.001670	0.003447	0.0065557
109	733.	0.001263	0.000340	0.006069	0.004647	0.015110	0.0005751
110	587.	0.000130	0.000209	0.000008	0.002056	0.000000	0.0004324
111	504.	0.000813	0.000247	0.001272	0.003540	0.004340	0.0424068
112	567.	0.000380	0.000021	0.000053	0.001165	0.000286	0.0001216

COLUMN COUNT	C1 112	COOK-B7 112	COOK-B8 112	COOK-B9 112	COOK-B10 112	COOK-B11 112
Row						
1	402.	0.006228	0.00288	0.00423	0.05285	0.000180
2	403.	0.001813	0.00428	0.12461	0.08587	0.026530
3	405.	0.000229	0.00003	0.00188	0.00135	0.006134
4	407.	0.000082	0.00023	0.00035	0.00016	0.000226
5	410.	0.000021	0.00002	0.00337	0.00814	0.000398
6	417.	0.000570	0.00852	0.00264	0.00017	0.001539
7	421.	0.001238	0.02210	0.00581	0.00100	0.004957
8	422.	0.000490	0.00837	0.00468	0.00001	0.001533
9	424.	0.000337	0.00001	0.00063	0.00031	0.000044
10	426.	0.000535	0.00003	0.00109	0.00028	0.000219
11	429.	0.000752	0.00029	0.00001	0.00005	0.001653
12	431.	0.000006	0.00002	0.00000	0.00007	0.000024
13	432.	0.012547	0.13211	0.01775	0.00822	0.031344
14	438.	0.000027	0.00714	0.00190	0.03177	0.000197
15	440.	0.000218	0.00007	0.05009	0.00682	0.035830
16	445.	0.000018	0.00009	0.00000	0.00159	0.001233
17	451.	0.002598	0.00001	0.00335	0.00483	0.000816
18	452.	0.000001	0.00006	0.00006	0.00036	0.000358
19	455.	0.000001	0.00000	0.00000	0.00005	0.000030
20	457.	0.000451	0.00250	0.00228	0.00445	0.006263
21	458.	0.000000	0.00000	0.00000	0.00000	0.000000
22	460.	0.000022	0.00016	0.00018	0.00005	0.000118
23	461.	0.002166	0.00457	0.00285	0.03979	0.020086
24	462.	0.000005	0.00008	0.00006	0.00014	0.000000
25	468.	0.001006	0.00288	0.00913	0.00483	0.000043
26	475.	0.000002	0.00014	0.00000	0.00081	0.000177
27	476.	0.000229	0.00019	0.00003	0.00038	0.000166
28	479.	0.000008	0.00024	0.00002	0.00005	0.000037
29	483.	0.000002	0.00053	0.00008	0.00106	0.002274
30	484.	0.016437	0.30481	0.00347	0.00340	0.009417
31	492.	0.000000	0.00001	0.00002	0.00016	0.000000
32	493.	0.003247	0.03741	0.00001	0.01140	0.011315
33	499.	0.013637	0.00000	0.00013	0.00009	0.001254
34	500.	0.000369	0.00184	0.00092	0.00032	0.002229
35	501.	0.001989	0.03245	0.04689	0.03830	0.000292
36	510.	0.013893	0.00017	0.00000	0.00229	0.013064
37	518.	0.000000	0.00000	0.00000	0.00000	0.000004
38	519.	0.002176	0.00009	0.00120	0.00022	0.003203
39	524.	0.000306	0.00025	0.00003	0.00384	0.002149
40	527.	0.000209	0.00125	0.00005	0.00012	0.000003
41	528.	0.000168	0.00174	0.00007	0.00059	0.000060
42	532.	0.003247	0.01001	0.00764	0.01564	0.004164
43	533.	0.000002	0.00028	0.00301	0.00885	0.003013
44	538.	0.000811	0.00211	0.00034	0.00000	0.003197
45	539.	0.000274	0.00019	0.00058	0.00172	0.000547
46	542.	0.001971	0.00290	0.00014	0.00109	0.001830
47	543.	0.002464	0.00003	0.02381	0.00367	0.012885
48	548.	0.025987	0.23118	0.02236	0.01887	0.001112
49	551.	0.000000	0.00002	0.00006	0.00002	0.000016
50	553.	0.000004	0.00026	0.00000	0.00021	0.000017
51	554.	0.001016	0.00071	0.00264	0.00467	0.001667
52	556.	0.000073	0.00006	0.00010	0.00040	0.000148
53	557.	0.000613	0.00001	0.02056	0.03662	0.000691
54	558.	0.000000	0.00021	0.00046	0.00040	0.000088
55	559.	0.000003	0.00002	0.00010	0.00006	0.000054
56	565.	0.000073	0.00049	0.00023	0.00000	0.000729

COLUMN COUNT	C1 112	COOK-B7 112	COOK-B8 112	COOK-B9 112	COOK-B10 112	COOK-B11 112
ROW						
57	568.	0.000005	0.00782	0.05252	0.04162	0.015401
58	571.	0.000012	0.01584	0.03558	0.00280	0.003886
59	575.	0.000007	0.00006	0.00001	0.00005	0.000043
60	577.	0.010522	0.01447	0.03420	0.01279	0.002211
61	581.	0.000019	0.00087	0.00045	0.00000	0.000275
62	582.	0.000082	0.00205	0.00204	0.00303	0.003400
63	583.	0.000409	0.00142	0.02466	0.07101	0.000010
64	585.	0.002097	0.01131	0.01165	0.00267	0.000004
65	465.	0.000106	0.00006	0.00087	0.00029	0.002046
66	544.	0.000389	0.00001	0.00005	0.00000	0.001315
67	478.	0.000000	0.00005	0.00003	0.00008	0.000650
68	443.	0.000092	0.00236	0.01001	0.00008	0.031659
69	526.	0.000014	0.00000	0.00001	0.00001	0.000000
70	485.	0.000034	0.00046	0.00048	0.00014	0.001605
71	489.	0.000933	0.00015	0.00088	0.00007	0.000664
72	525.	0.000986	0.00324	0.00405	0.00058	0.003466
73	464.	0.001300	0.00003	0.02060	0.01010	0.010548
74	436.	0.000241	0.00693	0.00188	0.00018	0.010333
75	563.	0.000495	0.00076	0.00003	0.00000	0.003179
76	430.	0.000009	0.00005	0.00002	0.00003	0.000108
77	413.	0.000016	0.00007	0.00012	0.00009	0.001520
78	420.	0.002443	0.00068	0.00022	0.00004	0.023137
79	579.	0.000952	0.00001	0.00012	0.00167	0.004083
80	419.	0.010575	0.00116	0.01623	0.00811	0.039535
81	586.	0.000027	0.00015	0.00002	0.00050	0.000248
82	590.	0.000821	0.03602	0.01516	0.01587	0.000525
83	406.	0.000543	0.00052	0.01190	0.01488	0.000049
84	418.	0.009604	0.00317	0.00961	0.01876	0.031743
85	423.	0.000047	0.00001	0.00011	0.00030	0.000020
86	434.	0.201470	7.72135	0.82787	0.02380	0.061797
87	441.	0.001680	0.10963	5.07969	2.54583	0.238163
88	788.	0.001249	0.00013	0.00003	0.00926	0.018886
89	645.	0.000279	0.01204	0.00203	0.00779	0.013644
90	447.	0.000299	0.01003	0.00032	0.00008	0.015991
91	448.	0.000183	0.00005	0.00002	0.00140	0.000936
92	449.	0.000031	0.00056	0.00163	0.00216	0.000483
93	454.	0.000002	0.00004	0.00004	0.00038	0.000296
94	459.	0.000497	0.00000	0.00003	0.00000	0.002180
95	654.	0.000797	0.00015	0.00003	0.00314	0.006827
96	661.	0.001807	0.00006	0.00113	0.00014	0.000224
97	736.	0.000001	0.00018	0.00030	0.00044	0.000220
98	502.	0.002416	0.01975	0.00006	0.00274	0.000111
99	512.	0.000075	0.00220	0.00083	0.00313	0.000442
100	514.	0.000000	0.00000	0.00000	0.00000	0.000000
101	516.	0.000428	0.00844	0.00001	0.00110	0.002441
102	776.	0.000008	0.00260	0.01403	0.00345	0.010342
103	894.	0.000073	0.00001	0.00113	0.00022	0.001791
104	692.	0.003656	0.00145	0.02311	0.00236	0.082170
105	535.	0.000227	0.00022	0.00019	0.00000	0.000001
106	541.	0.000218	0.00190	0.00926	0.00004	0.024169
107	574.	0.000760	0.00995	0.00042	0.00018	0.004504
108	580.	0.000288	0.00803	0.00001	0.00001	0.004055
109	733.	0.000333	0.00001	0.00049	0.01667	0.030223
110	587.	0.000148	0.00001	0.00007	0.00144	0.002388
111	504.	0.000016	0.00109	0.00032	0.00033	0.000382
112	567.	0.000108	0.00200	0.00019	0.00010	0.002339

APPENDIX F

LISTING OF $D_i(\beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11})$

APPENDIX F

LISTING OF $D_i(\beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11})$

COUNTER COUNT ROW	ID 112	DI(SUB) 112			
1	402.	0.02713	57	568.	0.01000
2	403.	0.11295	58	571.	0.00848
3	405.	0.00146	59	575.	0.00003
4	407.	0.00031	60	577.	0.00875
5	410.	0.00166	61	581.	0.00030
6	417.	0.00202	62	582.	0.00243
7	421.	0.00693	63	583.	0.01753
8	422.	0.00230	64	585.	0.00737
9	424.	0.00064	65	465.	0.00062
10	426.	0.00089	66	544.	0.00033
11	429.	0.00085	67	478.	0.00024
12	431.	0.00005	68	443.	0.01257
13	432.	0.03708	69	526.	0.00001
14	438.	0.01066	70	485.	0.00068
15	440.	0.04128	71	489.	0.00048
16	445.	0.00059	72	525.	0.00312
17	451.	0.00175	73	464.	0.00996
18	452.	0.00013	74	436.	0.00598
19	455.	0.00002	75	563.	0.00086
20	457.	0.00198	76	430.	0.00006
21	458.	0.00000	77	413.	0.00032
22	460.	0.00027	78	420.	0.00558
23	461.	0.01912	79	579.	0.00196
24	462.	0.00010	80	419.	0.00986
25	468.	0.00454	81	586.	0.00035
26	475.	0.00036	82	590.	0.01072
27	476.	0.00036	83	406.	0.00478
28	479.	0.00008	84	418.	0.00991
29	483.	0.00072	85	423.	0.00022
30	484.	0.09207	86	434.	1.61840
31	492.	0.00004	87	441.	1.14773
32	493.	0.01239	88	788.	0.00682
33	499.	0.00278	89	645.	0.00633
34	500.	0.00155	90	447.	0.00747
35	501.	0.01395	91	448.	0.00069
36	510.	0.00860	92	449.	0.00084
37	518.	0.00000	93	454.	0.00013
38	519.	0.00171	94	459.	0.00046
39	524.	0.00191	95	654.	0.00285
40	527.	0.00032	96	661.	0.00117
41	528.	0.00060	97	736.	0.00012
42	532.	0.00568	98	502.	0.00664
43	533.	0.00187	99	512.	0.00146
44	538.	0.00161	100	514.	0.00000
45	539.	0.00125	101	516.	0.00351
46	542.	0.00252	102	776.	0.00420
47	543.	0.01346	103	894.	0.00055
48	548.	0.08369	104	692.	0.03071
49	551.	0.00003	105	535.	0.00010
50	555.	0.00017	106	541.	0.01047
51	554.	0.00277	107	574.	0.00409
52	556.	0.00023	108	580.	0.00290
53	557.	0.00871	109	733.	0.01140
54	558.	0.00044	110	587.	0.00067
55	559.	0.00010	111	504.	0.00030
56	565.	0.00025	112	567.	0.00086

APPENDIX G

QR DECOMPOSITION

APPENDIX G

QR DECOMPOSITION

Due to problems with round-off error, computer storage, and efficiency of other algorithms, most regression computer programs do not solve the normal equations as such to find the least squares solutions. Instead, many programs rely on techniques based on the QR decomposition (or something similar) of the X matrix. After a brief description of this method, we will indicate how the necessary information can be obtained to calculate the desired diagnostics. Let X be an $(n \times p)$ matrix of rank p . Then there exists an orthogonal matrix Q ($n \times n$) such that

$$Q'X = \begin{bmatrix} R \\ 0 \end{bmatrix} \quad \left. \begin{array}{l} \text{ } (p \times p) \\ \text{ } (n-p) \times p \end{array} \right\}$$

where R is an upper triangular matrix. If we partition $Q = [Q_1 \quad Q_2]$ where Q_1 is an $(n \times p)$ matrix, we have

$$Q_1'Q_1 = I_p, Q_1'Q_2 = 0, Q_2'Q_2 = I_{n-p},$$

and $X = Q_1R$. This last representation is called the QR decomposition of X .

In a regression situation where $\hat{X}\underline{\hat{\beta}} = Y$, this decomposition yields

$$\underline{\hat{\beta}} = R^{-1}Q_1'Y$$

as the solution. However, the equations $R\underline{\hat{\beta}} = Q_1'Y$ are usually solved by back-substitution, taking advantage of the fact that R is an upper triangular matrix. As mentioned in Chapter 3, the principal information needed for regression diagnostics are the vector of residuals ($\underline{r}$), the vector of values on the diagonal of the hat matrix ($\underline{h}$), the value of s , and the matrix C . Using the properties of Q_1 and R , the following results are straightforward:

$$a) \quad \underline{r} = (I - Q_1Q_1')Y = Q_2Q_2'Y$$

$$b) \quad H = X(X'X)^{-1}X' = Q_1RR^{-1}(R')^{-1}R'Q_1' = Q_1Q_1'$$

$$c) \quad s^2 = \left(\frac{1}{n-p}\right) \underline{Y}'(I - Q_1 Q_1') \underline{Y} = (Q_2' \underline{Y})' Q_2' \underline{Y}$$

$$d) \quad C = (X'X)^{-1} X' = R^{-1} Q_1'$$

These relationships will give the analyst all the information necessary to compute regression diagnostics. For other information on QR decomposition and its use in least-squares regression, see the text by Seber (1977, pp. 308-316). The book by Stewart (1973) contains general discussion about QR decomposition, and the Linpack User's Guide (1979, Chapter 9) describes programs to implement QR decomposition. In fact, the subroutines SQRDC, SQRSL, and STRSL can be used to calculate R , $\hat{\beta}$, and h_i , respectively. Belsley, Kuh and Welsch (1980, pp. 69-75) and Cook and Weisberg (1980, p. 502) discuss computational considerations for the calculation of regression diagnostics.

DISTRIBUTION

No. of
Copies

OFFSITE

A. A. Churm
DOE Patent Division
9800 S. Cass Avenue
Argonne, IL 60439

D. M. Austin
DOE Division of Engineering,
Mathematical and Geosciences
Washington, DC 20545

27 DOE Technical Information Center

A. Shoshani
Computer Science and
Applied Mathematics Dept.
Lawrence Berkeley Laboratory
Berkeley, CA 94720

R. W. Mensing
Numerical Mathematics Group
Lawrence Livermore Laboratory
P.O. Box 808
Livermore, CA 94550

R. A. Waller
Group Q-12, MS-608
Los Alamos National Laboratory
Los Alamos, NM 87545

P. Bloomfield
Department of Statistics
Princeton University, Fine Hall
Princeton, NJ 08540

J. W. Wilkinson
School of Management
Rensselaer Polytechnic Institute
Troy, NY 12181

J. H. Friedman
Stanford Linear Accelerator Center
P.O. Box 4349
Stanford, CA 94305

No. of
Copies

OFFSITE

J. Heller
Department of Computer Science
SUNY at Stony Brook
Stony Brook, NY 11794

W. Dixon
Department of Mathematics
UCLA Medical Center
Los Angeles, CA 90024

V. E. Kane
Mathematics & Statistics
Research Dept.
Union Carbide Corp. - Nuclear Div.
P.O. Box Y
Oak Ridge, TN 37830

D. L. Wallace
Department of Statistics
University of Chicago
1118 E. 58th Street
Chicago, IL 60637

D. J. DeWitt
Computer Sciences Dept.
University of Wisconsin
Madison, WI 53706

B. J. Joiner
Statistics Department
University of Wisconsin
Madison, WI 53706

B. J. Tepping
401 Apple Grove Road
Silver Springs, MD 20904

DISTRIBUTION

No. of
Copies

ONSITE

DOE Richland Operations Office

H. E. Ransom

32 Pacific Northwest Laboratory

R. A. Burnett
D. B. Carr
J. C. Daly
M. D. Erickson
D. L. Hall (3)
R. J. Littlefield
A. R. Olsen (17)
W. L. Nicholson
J. J. Thomas
Publishing Coordination (2)
Technical Information (3)

DO NOT
MICROFILM