The Hat Matrix in Regression and ANOVA

DAVID C. HOAGLIN AND ROY E. WELSCH*

In least-squares fitting it is important to understand the influence which a data y value will have on each fitted y value. A projection matrix known as the hat matrix contains this information and, together with the Studentized residuals, provides a means of identifying exceptional data points. This approach also simplifies the calculations involved in removing a data point, and it requires only simple modifications in the preferred numerical least-squares algorithms.

KEY WORDS: Analysis of variance; Regression analysis; Projection matrix; Outliers; Studentized residuals; Least-squares computations.

1. Introduction

In fitting linear models by least squares it is very often useful to determine how much influence or leverage each data y value (y_j) can have on each fitted y value ($\hat{y}_i$). For the fitted value $\hat{y}_i$ corresponding to the data value y_i, the relationship is particularly straightforward to interpret, and it can reveal multivariate outliers among the carriers (or x variables) which might otherwise be difficult to detect. The desired information is available in the hat matrix, which gives each fitted value $\hat{y}_i$ as a linear combination of the observed values y_j. (The term "hat matrix" is due to John W. Tukey, who introduced us to the technique about ten years ago.) The present article derives and discusses the hat matrix and gives an example to illustrate its usefulness.

Section 2 defines the hat matrix and derives its basic properties. Section 3 formally examines two familiar examples, while Section 4 gives a numerical example. In practice one must, of course, consider the actual effect of the data y values in addition to their leverage; we discuss this in terms of the residuals in Section 5. Section 6 then sketches how the hat matrix can be obtained from two accurate numerical algorithms used for solving least-squares problems.

2. Basic Properties

We are concerned with the linear model

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{X}\ \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}\,, \tag{2.1}$$

which summarizes the dependence of the *response y* on the *carriers* $X_1, \ldots, X_p$ in terms of the data values y_i and $x_{i1}, \ldots, x_{ip}$ for $i = 1, \ldots, n$. (We refrain from thinking of $X_1, \ldots, X_p$ as independent variables because they are often not independent in any reasonable sense.) In fitting the model (2.1) by least squares (assuming that X has rank p and that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2 \mathrm{I}_n$), we usually obtain the fitted or predicted values from $\hat{\mathbf{y}} = X\mathbf{b}$, where $\mathbf{b} = (X^TX)^{-1}X^T\mathbf{y}$. From this it is simple to see that

$$\hat{\mathbf{y}} = X(X^TX)^{-1}X^T\mathbf{y}. \tag{2.2}$$

To emphasize the fact that (when X is fixed) each $\hat{y}_i$ is a linear function of the y_j, we write (2.2) as

$$\hat{\mathbf{y}} = H\mathbf{y}, \tag{2.3}$$

where $H = X(X^TX)^{-1}X^T$. The $n \times n$ matrix H is known as the hat matrix simply because it maps $\mathbf{y}$ into $\hat{\mathbf{y}}$. Geometrically, if we represent the data vector $\mathbf{y}$ and the columns of X as points in euclidean n space, then the points $X\boldsymbol{\beta}$ (which we can obtain as linear combinations of the column vectors) constitute a p dimensional subspace. The fitted vector $\hat{\mathbf{y}}$ is the point of that subspace nearest to $\mathbf{y}$, and it is also the perpendicular projection of $\mathbf{y}$ into the subspace. Thus H is a projection matrix. Also familiar is the role which H plays in the covariance matrices of $\hat{\mathbf{y}}$ and of $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$:

$$\mathrm{var}(\hat{\mathbf{y}}) = \sigma^2 H, \tag{2.4}$$

$$\mathrm{var}(\mathbf{r}) = \sigma^2(I - H). \tag{2.5}$$

For the data analyst, the element h_{ij} of H has a direct interpretation as the amount of leverage or influence exerted on $\hat{y}_i$ by y_j (regardless of the actual value of y_j, since H depends only on X). Thus a look at the hat matrix can reveal sensitive points in the design, points at which the value of y has a large impact on the fit (Huber 1975). In using the word "design" here, we have in mind both the standard regression or ANOVA situation, in which the values of $X_1, \ldots, X_p$ are fixed in advance, and the situation in which y and $X_1, \ldots, X_p$ are sampled together. The simple designs, such as two-way analysis of variance, give good control over leverage (as we shall see in Section 3); and with fixed X one can examine, and perhaps modify, the experimental conditions in advance. When the carriers are sampled, one can at least determine whether the observed X contains sensitive points and consider omitting them if the corresponding y value seems discrepant. Thus we use the hat matrix to identify "high-leverage points." If this notion is to be really useful, we must make it more precise.

The influence of the response value y_i on the fit is most directly reflected in its leverage on the corre-

* David C. Hoaglin is Senior Analyst, Abt Associates, 55 Wheeler Street, Cambridge, MA 02138, and Research Associate, Department of Statistics, Harvard University, Cambridge, MA 02138. Roy E. Welsch is Associate Professor of Operations Research and Management, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139. This work was supported in part by NSF Grant SOC75-15702 to Harvard University and by NSF Grant 76-14311 DSS to the National Bureau of Economic Research.

sponding fitted value $\hat{y}_i$, and this is precisely the information contained in h_{ii}, the corresponding diagonal element of the hat matrix. We can easily imagine fitting a simple regression line to data (x_i, y_i), making large changes in the y value corresponding to the largest x value, and watching the fitted line follow that data point. In this one-carrier problem or in a two-carrier problem a scatter plot will quickly reveal any x outliers, and we can verify that they have relatively large diagonal elements h_{ii}. When $p > 2$, scatter plots may not reveal multivariate outliers, which are separated in p space from the bulk of the x points but do not appear as outliers in a plot of any single carrier or pair of carriers, and the diagonal of the hat matrix is a source of valuable diagnostic information. In addition to being somewhat easier to understand, the diagonal elements of H can be less trouble to compute, store, and examine, especially if n is moderately large. Thus attention focuses primarily (often exclusively) on the h_{ii}, which we shall sometimes abbreviate h_i. We next examine some of their properties.

As a projection matrix, H is symmetric and idempotent ($H^2 = H$), as we can easily verify from the definition following (2.3). Thus we can write

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2, \tag{2.6}$$

and it is immediately clear that $0 \leq h_{ii} \leq 1$. These limits are helpful in understanding and interpreting h_{ii}, but they do not yet tell us when h_{ii} is large. We know, however, that the eigenvalues of a projection matrix are either zero or one and that the number of nonzero eigenvalues is equal to the rank of the matrix. In this case, $\text{rank}(H) = \text{rank}(X) = p$, and hence $\text{trace}(H) = p$, i.e.,

$$\sum_{i=1}^{n} h_i = p. \tag{2.7}$$

The average size of a diagonal element of the hat matrix, then, is p/n. Experience suggests that a reasonable rule of thumb for large h_i is $h_i > 2p/n$. Thus we determine high-leverage points by looking at the diagonal elements of H and paying particular attention to any x point for which $h_i > 2p/n$. Usually we treat the n values h_i as a batch of numbers and bring them together in a stem-and-leaf display (as we shall illustrate in Section 4). For a more refined screening when the model includes the constant carrier and the rows of X are sampled from a $(p - 1)$ variate Gaussian distribution, we could use the fact that (for any single h_i) $[(n - p)(h_i - 1/n)]/[(p - 1)(1 - h_i)]$ has an F distribution on $p - 1$ and $n - p$ degrees of freedom.

From (2.6), we can also see that whenever $h_{ii} = 0$ or $h_{ii} = 1$, we have $h_{ij} = 0$ for all $j \neq i$. These two extreme cases can be interpreted as follows. First, if $h_{ii} = 0$, then $\hat{y}_i$ must be fixed at zero by design—it is not affected by y_i or by any other y_j. A point with $x = 0$ when the model is a straight line through the origin provides a simple example. Second, when $h_{ii} = 1$, we have $\hat{y}_i = y_i$—the model always fits this data value exactly. In effect, the model dedicates a parameter to this particular observation (as is sometimes done explicitly by adding a dummy variable to remove an outlier).

Now that we have developed the hat matrix and a number of its properties, we turn to three examples, two designed and one sampled. We then discuss (in Section 5) how to handle y_i when h_{ii} indicates a high-leverage point.

3. Formal Examples

To illustrate the hat matrix and develop our intuition, we begin with two familiar examples in which the calculations can be done by simple algebra.

The usual regression line,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

has

$$X = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}^T,$$

and a few steps of algebra give

$$h_{ij} = \frac{1}{n} + [(x_i - \bar{x})(x_j - \bar{x})] \bigg/ \left[\sum_{k=1}^{n} (x_k - \bar{x})^2\right]. \tag{3.1}$$

Next we examine the relationship between structure and leverage in a simple balanced design: a two-way table with R rows and C columns and one observation per cell. (Behnken and Draper (1972) discuss variances of residuals in several more complicated designs. It is straightforward to find H through (2.5).) The usual model for the $R \times C$ table is

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

with the constraints $\alpha_1 + \ldots + \alpha_R = 0$ and $\beta_1 + \ldots + \beta_C = 0$; here $n = RC$ and $p = R + C - 1$. We could, of course, write this model in the form of (2.1), but it is simpler to preserve the subscripts i and j and to denote an element of the hat matrix as $h_{ij,kl}$. When we recall that

$$\hat{y}_{ij} = y_{i\cdot} + y_{\cdot j} - y_{\cdot\cdot} \tag{3.2}$$

(a dot in place of a subscript indicates the average with respect to that subscript), it is straightforward to obtain

$$h_{ij,ij} = 1/C + (1/R) - (1/RC) = (R + C - 1)/RC; \tag{3.3}$$

$$h_{ij,il} = (R - 1)/RC, \qquad l \neq j; \tag{3.4}$$

$$h_{ij,kj} = (C - 1)/RC, \qquad k \neq i; \tag{3.5}$$

$$h_{ij,kl} = -(1/RC), \qquad k \neq i, l \neq j. \tag{3.6}$$

From (3.3) we see that all the diagonal elements of H are equal, as we would expect in a balanced design. Further, (3.3) through (3.6) show that $\hat{y}_{ij}$ will be affected by any change in y_{kl} for any values of k and l.

4. A Numerical Example

In this section we examine the hat matrix in a regression example, emphasizing (either here or in Section 5) the connections between it and other sources of diagnostic information. We use a ten-point example, for which we can easily present H in full. In a larger data set, we would generally work with only the diagonal elements, h_i. Welsch and Kuh (1977) discuss a larger example.

The data for this example come from Draper and Stoneman (1966); we reproduce it in Table 1. The response is strength, and the carriers are the constant, specific gravity, and moisture content. To probe the relationship between the nonconstant carriers, we plot moisture content against specific gravity (Figure A). In this plot, point 4, with coordinates (0.441, 8.9), is to some extent a bivariate outlier (its value is not extreme for either carrier), and we should expect it to have substantial leverage on the fit. Indeed, if this point were absent, it would be considerably more difficult to distinguish the two carriers.

1. Data on Wood Beams

beam number	specific gravity	moisture content	strength
1	0.499	11.1	11.14
2	0.558	8.9	12.74
3	0.604	8.8	13.13
4	0.441	8.9	11.51
5	0.550	8.8	12.38
6	0.528	9.9	12.60
7	0.418	10.7	11.13
8	0.480	10.5	11.70
9	0.406	10.5	11.02
10	0.467	10.7	11.41

The hat matrix for this X appears in Table 2, and a stem-and-leaf display (Tukey 1972b, 1977) of the diagonal elements (rounded to multiples of .01) is as follows:

0	
1	559
2	456
3	2
4	22
5	
6	0

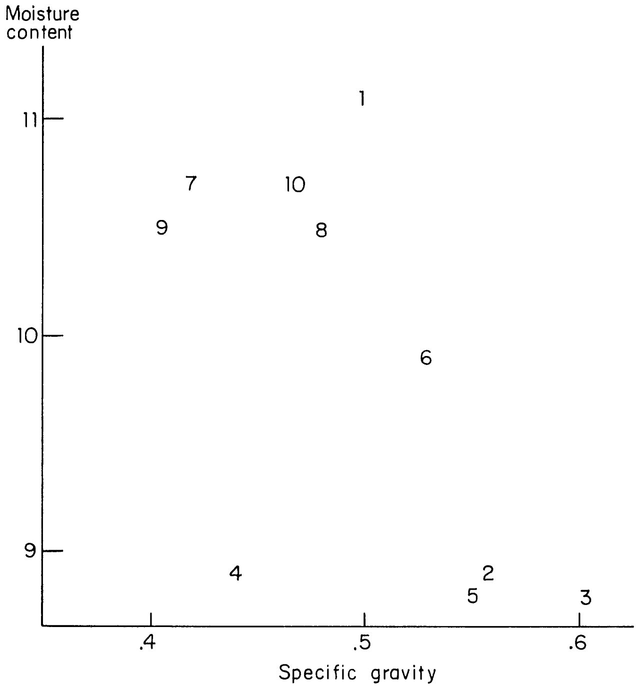

Figure A. The Two Carriers for the Wood Beam Data (Plotting symbol is beam number.).

We note that h_4 is the largest diagonal element and that it just exceeds the level $(2p/n = 6/10)$ set by our rough rule of thumb. Examining H element by element, we find that it responds to the other qualitative features of Figure A. For example, the relatively high leverage of points 1 and 3 reflects their position as extremes in the scatter of points. The moderate negative value of $h_{1,4}$ is explained by the positions of points 1 and 4 on opposite sides of the rough sloping band where the rest of the points lie. The moderate positive values of $h_{1,8}$ and $h_{1,10}$ show the mutually reinforcing positions of these three points. The central position of point 6 accounts for its low leverage. Other noticeable values of h_{ij} have similar explanations.

Having identified point 4 as a high-leverage point in this data set, it remains to investigate the effect of its position and response value on the fit. Does the model fit well at point 4, or should this point be set aside? We turn to these questions next.

2. The Hat Matrix for the Wood Beam Data (lower triangle omitted by symmetry)

	j									
i	1	2	3	4	5	6	7	8	9	10
1	.418	-.002	.079	-.274	-.046	.181	.128	.222	.050	.242
2		.242	.292	.136	.243	.128	-.041	.033	-.035	.004
3			.417	-.019	.273	.187	-.126	.044	-.153	.004
4				.604	.197	-.038	.168	-.022	.275	-.028
5					.252	.111	-.030	.019	-.010	-.010
6						.148	.042	.117	.012	.111
7							.262	.145	.277	.174
8								.154	.120	.168
9									.315	.148
10										.187

5. Bringing in the Residuals

So far we have examined the design matrix X for evidence of points where the data value y has high leverage on the fitted value $\hat{y}$. If such influential points are present, we must still determine whether they have had any adverse effects on the fit. A discrepant value of y, especially at an influential design point, may lead us to set that entire observation aside (planning to investigate it in detail separately) and refit without it, but we emphasize that such decisions cannot be made automatically. As we can see for the regression line, with h_{ij} given by (3.1), the more extreme design points generally provide the greatest information on certain coefficients (in this case, the slope), and omitting such an observation may substantially reduce the precision with which we can estimate those coefficients. If we delete row i, that is, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, from the design matrix X and denote the result by $X_{(i)}$, then (Rao 1965, p. 29), except for the constant factor σ^2, the covariance matrix of $\mathbf{b}$ is

$$(X_{(i)}^T X_{(i)})^{-1} = (X^T X)^{-1} + (X^T X)^{-1}\mathbf{x}_i{}^T\mathbf{x}_i(X^T X)^{-1}/(1 - h_i). \qquad (5.1)$$

The presence of $(1 - h_i)$ in the denominator shows how removing a high-leverage point may increase the variance of coefficient estimates. Alternatively, the accuracy of the apparently discrepant point may be beyond question, so that dismissing it as an outlier would be unacceptable. In both these situations, then, the apparently discrepant point may force us to question the adequacy of the model.

In detecting discrepant y values, we always examine the residuals, $r_i = y_i - \hat{y}_i$, using such techniques as a scatterplot against each carrier, a scatterplot against $\hat{y}$, and a normal probability plot. (Anscombe (1973) has discussed and illustrated some of these.) When there is substantial variation among the h_i values, (2.5) indicates that we should allow for differences in the variances of the r_i (Anscombe and Tukey 1963) and look at $r_i/(1 - h_i)^{1/2}$. This adjustment puts the residuals on an equal footing, but it is often more convenient to use the *standardized residual,* $r_i/(s(1 - h_i)^{1/2})$, where s^2 is the residual mean square.

For diagnostic purposes, we would naturally ask about the size of the residual corresponding to y_i when data point i has been omitted from the fit. That is, we base the fit on the remaining $n - 1$ data points and then predict the value for y_i. This residual is $y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{(i)}$, where $\hat{\boldsymbol{\beta}}_{(i)}$ is the least-squares estimate of $\boldsymbol{\beta}$ based on all the data except data point i. (These residuals are also the basis of Allen's (1974) PRESS criterion for selecting variables in regression.) Similarly $s_{(i)}^2$ is the residual mean square for the "not-i" fit, and the standard deviation of $y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{(i)}$ is estimated by $s_{(i)}[1 + \mathbf{x}_i(X_{(i)}^T X_{(i)})^{-1}\mathbf{x}_i{}^T]^{1/2}$. We now define the *Studentized residual*:

$$r_i^* = \frac{y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{(i)}}{s_{(i)}[1 + \mathbf{x}_i(X_{(i)}^T X_{(i)})^{-1}\mathbf{x}_i{}^T]^{1/2}}. \qquad (5.2)$$

Since the numerator and denominator in (5.2) are independent, r_i^* has a t distribution on $n - p - 1$ degrees of freedom, and we can readily assess the significance of any single Studentized residual. (Of course, r_i^* and r_j^* will not be independent.) In actually calculating the Studentized residuals we can save a great deal of effort by observing that the quantities we need are readily available. Straightforward algebra using (5.1) turns (5.2) into

$$r_i^* = r_i/(s_{(i)}(1 - h_i)^{1/2}), \qquad (5.3)$$

and we can obtain $s_{(i)}$ from

$$(n - p - 1)s_{(i)}^2 = (n - p)s^2 - r_i{}^2/(1 - h_i). \qquad (5.4)$$

Once we have the diagonal elements of H, the rest is simple.

Our diagnostic strategy, then, is to examine the h_i for high-leverage design points and the r_i^* for discrepant y values. These two aspects of the search for troublesome data points are complementary; neither is sufficient by itself. When h_i is small, r_i^* may be large because r_i is large, but the impact of y_i on the fit or on the coefficients may be minor. And when h_i is large, r_i^* may still be moderate or small because y_i is consistent with the model and the rest of the data.

Just how to combine the information from h_i and r_i^* is a matter of judgment. We prefer the more detailed grasp of the data which comes from looking at the h_i and the r_i^* separately. For diagnostic purposes, a practice which we recommend is to tag as exceptional any data point for which h_i or r_i^* is significant at the 10 percent level. To decide whether an exceptional point is actually damaging, one would then use a criterion which is appropriate in the context of the data. Two likely criteria are the change in coefficients, $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$, easily calculated from

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{(i)} = (X^T X)^{-1}\mathbf{x}_i{}^T r_i/(1 - h_i); \qquad (5.5)$$

and the change in fit at point i, $\mathbf{x}_i(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$, which simply reduces to $h_i r_i/(1 - h_i)$. (The size of such changes would customarily be compared to some suitable measure of scale.) For both of these criteria it is easy to determine the effect of setting aside an exceptional point without recalculation.

To continue our diagnosis of the wood beam example, we plot strength against specific gravity in Figure B and strength against moisture content in Figure C. With the exception of beam 1, the first of these looks quite linear and well-behaved. In the second plot we see somewhat more scatter, and beam 4 stands apart from the rest. Table 3 gives r_i, $(1 - h_i)^{1/2}$, $s_{(i)}$, and the Studentized residuals r_i^*. Among the r_i^*, beam 1 appears as a clear stray ($p < .02$), and beam 6 also deserves attention ($p < .1$). Since beam 4 is known to have high leverage ($h_i = .604$), we continue to investigate it.

The fit for the full data is

$$\hat{y} = 10.302 + 8.495(SG) - 0.2663(MC), \qquad (5.6)$$

with $s = 0.2753$; and when we set aside beams 1, 4,

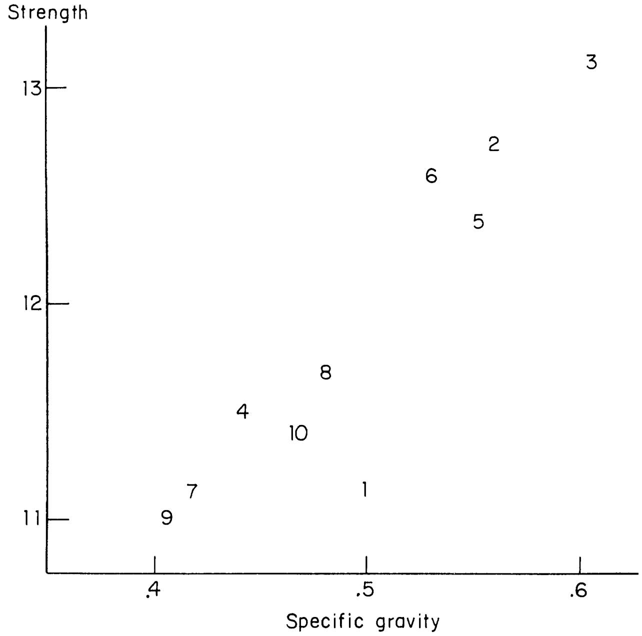

Figure B. Strength versus Specific Gravity for the Wood Beam Data (Plotting symbol is beam number.).

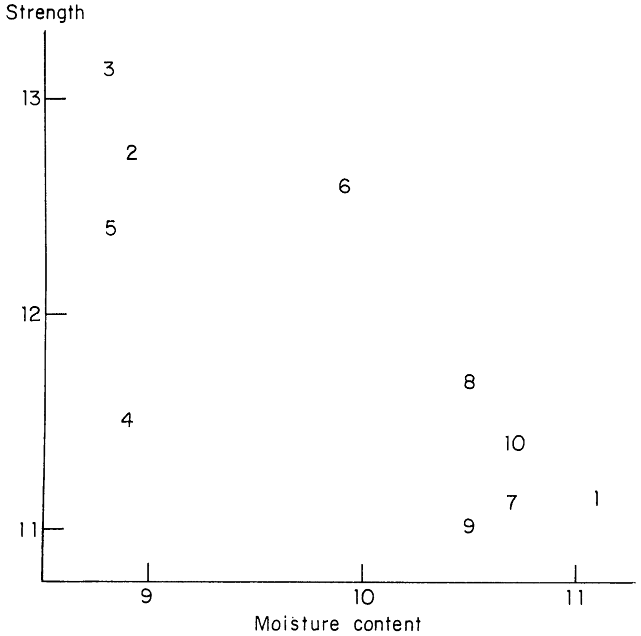

Figure C. Strength versus Moisture Content for the Wood Beam Data (Plotting symbol is beam number.).

and 6 in turn, we find $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$ to be $(2.710, -1.772, -0.1932)^T$, $(-2.109, 1.695, 0.1242)^T$, and $(-0.642, 0.748, 0.0329)^T$, respectively. The estimated standard errors for $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are 1.896, 1.784, and 0.1237, so that setting aside either beam 1 or beam 4 causes each coefficient to change by roughly 1.0 to 1.5 in standard-error units. Thus we should be reluctant to include these data points. By comparison, removing beam 6 leads to changes only about 25 percent as large.

Similarly, the change in fit at point i, $\mathbf{x}_i(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$, is -0.319 for beam 1, -0.256 for beam 4, and 0.078 for beam 6. Dividing each of these by the estimated standard error of $\hat{y}_i$ ($s\sqrt{h_i}$ from (2.4)) yields -1.790, -1.196, and 0.737, respectively. On the whole these are not as substantial as the coefficient changes, but beam 1 and (to a lesser extent) beam 4 are still fairly damaging.

3. Studentized Residuals and Related Quantities for the Wood Beam Data

i	r_i	h_i	$(1-h_i)^{1/2}$	$s_{(i)}$	r_i^*
1	-.444	.418	.763	.179	-3.254
2	.069	.242	.871	.296	.267
3	.041	.417	.764	.297	.182
4	-.167	.604	.629	.277	-.961
5	-.250	.252	.865	.273	-1.058
6	.450	.148	.923	.221	2.203
7	.127	.262	.859	.291	.509
8	.117	.154	.920	.293	.436
9	.066	.315	.828	.296	.270
10	-.009	.187	.902	.298	-.033

We have used two sources of diagnostic information, the diagonal elements of the hat matrix and the Studentized residuals, to identify data points which may have an unusual impact on the results of fitting the linear model (2.1) by least squares. We must interpret this information as clues to be followed up to determine whether a particular data point is discrepant, but not as automatic guidance for discarding observations. Often the circumstances surrounding the data will provide explanations for unusual behavior, and we will be able to reach a much more insightful analysis. Judgment and external sources of information can be important at many stages. For example, if we were trying to decide whether to include moisture content in the model for the wood beam data (the context in which Draper and Stoneman (1966) introduced this example), we would have to give close attention to the effect of beam 4 on the correlation between the carriers as well as the correlation between the coefficients. Such considerations do not readily lend themselves to automation and are an important ingredient in the difference between data analysis and data processing (Tukey 1972a).

6. Computation

Since we find the hat matrix (at least the diagonal elements h_i) a very worthwhile diagnostic addition to the information usually available in multiple regression, we now briefly describe how to obtain H from the more accurate numerical techniques for solving least-squares problems. Just as these techniques provide greater accuracy by not forming X^TX or solving the normal equations directly, we do not calculate H according to the definition.

For most purposes the method of choice is to represent X as

$$\underset{n\times p}{X} = \underset{n\times n}{Q}\ \underset{n\times p}{R}, \tag{6.1}$$

(with Q an orthogonal transformation and $R = [\tilde{R}^T,$

$0^T]^T$, where $\tilde{R}$ is $p \times p$ upper triangular) and obtain Q as a product of Householder transformations. Substituting (6.1) and the special structure of R into the definition of H, we see that

$$H = Q\begin{bmatrix} I_p & 0 \\ 0 & 0 \end{bmatrix}Q^T. \tag{6.2}$$

With a modest increase in computation cost, a simple modification of the basic algorithm yields H as a byproduct. If n is large, we can arrange to calculate and store only the h_i.

Finally we mention the singular-value decomposition,

$$\underset{n\times p}{X} = \underset{n\times p}{U}\;\underset{p\times p}{\Sigma}\;\underset{p\times p}{V^T}, \tag{6.3}$$

where $U^TU = I_p$, Σ is diagonal, and V is orthogonal. If this more elaborate approach is used (e.g., when X might not be of full rank), we can calculate the hat matrix from

$$H = UU^T. \tag{6.4}$$

These and other decompositions are discussed by Golub (1969). For a recent account of numerical techniques in solving linear least-squares problems, we recommend the book by Lawson and Hanson (1974).

[*Received October 18, 1976. Revised June 9, 1977.*]

References

Allen, D. M. (1974), "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics,* 16, 125–127.

Anscombe, F. J. (1973), "Graphs in Statistical Analysis," *The American Statistician,* 27, 17–21.

———, and Tukey, J. W. (1963), "The Examination and Analysis of Residuals," *Technometrics,* 5, 141–160.

Behnken, D. W., and Draper, N. R. (1972), "Residuals and Their Variance Patterns," *Technometrics,* 14, 101–111.

Draper, N. R., and Stoneman, D. M. (1966), "Testing for the Inclusion of Variables in Linear Regression by a Randomisation Technique," *Technometrics,* 8, 695–699.

Golub, G. H. (1969), "Matrix Decompositions and Statistical Calculations," in *Statistical Computation,* eds. R. C. Milton and J. A. Nelder, New York: Academic Press.

Huber, P. J. (1975), "Robustness and Designs," in *A Survey of Statistical Design and Linear Models,* ed. J. N. Srivastava, Amsterdam: North-Holland Publishing Co.

Lawson, C. L., and Hanson, R. J. (1974), *Solving Least Squares Problems,* Englewood Cliffs, N.J.: Prentice-Hall.

Rao, C. R. (1965), *Linear Statistical Inference and Its Applications,* New York: John Wiley & Sons.

Tukey, J. W. (1972a), "Data Analysis, Computation and Mathematics," *Quarterly of Applied Mathematics,* 30, 51–65.

——— (1972b), "Some Graphic and Semigraphic Displays," in *Statistical Papers in Honor of George W. Snedecor,* ed. T. A. Bancroft, Ames, Iowa: Iowa State University Press.

——— (1977), *Exploratory Data Analysis,* Reading, Mass.: Addison-Wesley Publishing Co.

Welsch, R. E., and Kuh, E. (1977), "Linear Regression Diagnostics," Working Paper 173, Cambridge, Mass.: National Bureau of Economic Research.