THE HAT MATRIX IN REGRESSION AND ANOVA

David C. Hoaglin, Harvard University*
Roy E. Welsch, Massachusetts Institute of Technology
and National Bureau of Economic
Research**

WP 904-77 January 1977

* Supported in part by NSF grant SOC75-15702 to Harvard University

** Supported in part by NSF grant 76-14311 DSS to the National Bureau of Economic Research

The Hat Matrix in Regression and ANOVA

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_i) can have on each fitted y-value ($\hat{y}_j$). 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. In a regression problem 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 paper derives and discusses the hat matrix and gives several examples which illustrate its usefulness.

Section 2 defines the hat matrix and derives its basic properties. Section 3 formally examines some familiar simple examples, while Section 4 gives two numerical examples. 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 some of the numerical algorithms used for solving least-squares problems.

2. BASIC PROPERTIES

We are concerned with the linear model

$$\underset{n\times 1}{\underset{\sim}{y}} = \underset{n\times p}{X}\ \underset{p\times 1}{\underset{\sim}{\beta}} + \underset{n\times 1}{\underset{\sim}{\varepsilon}} \quad , \tag{2.1}$$

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 $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$ and $\mathrm{var}(\underset{\sim}{\varepsilon}) = \sigma^2 I_n$), we usually obtain the fitted or predicted values from $\hat{\underset{\sim}{y}} = X\underset{\sim}{b}$, where $\underset{\sim}{b} = (X^TX)^{-1}X^T\underset{\sim}{y}$. From this it is simple to see that

$$\hat{\underset{\sim}{y}} = X(X^TX)^{-1}X^T\,\underset{\sim}{y} \quad . \tag{2.2}$$

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

$$\hat{\underset{\sim}{y}} = H\underset{\sim}{y} \quad , \tag{2.3}$$

where $H = X(X^TX)^{-1}X^T$. The $n\times n$ matrix H is known as "the hat matrix" simply because it takes $\underset{\sim}{y}$ into $\hat{\underset{\sim}{y}}$. Geometrically

$\hat{\underset{\sim}{y}}$ is the projection of $\underset{\sim}{y}$ onto the p-dimensional subspace of n-space spanned by the columns of X. Also familiar is the role which H plays in the covariance matrices of $\hat{\underset{\sim}{y}}$ and of $\underset{\sim}{r} = \underset{\sim}{y} - \hat{\underset{\sim}{y}}$:

$$\text{var}(\hat{\underset{\sim}{y}}) = \sigma^2 H \qquad (2.4)$$

$$\text{var}(\underset{\sim}{r}) = \sigma^2 (I-H) \quad . \qquad (2.5)$$

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 [7]. 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 corresponing 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 corresponding 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 the 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 below (2.3). Thus we can write

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \quad , \qquad (2.6)$$

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". It is easy to show, however, that the eigenvalues of a projection matrix are either 0 or 1 and that the number of non-zero eigenvalues is equal to the rank of the matrix. In this case, rank(H) = rank(X) = p , and hence trace(H) = p , that is,

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

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 h_i values as a batch of numbers and bring them together in a stem-and-leaf display (as we shall illustrate in Section 4).

From equation (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. We examine this situation further in the appendix.

Now that we have developed the hat matrix and a number of its properties, we turn to a variety of examples, some designed and some 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 use of the hat matrix and develop our intuition, we begin with a few familiar examples in which the calculations can be done by simple algebra. The most basic of these is the sample mean: $\hat{y}_i = \bar{y}$ for all i, and every element of H is $1/n$. Here $p=1$, and each $h_i=p/n$.

For a straight line through the origin, $X = (x_1,\ldots,x_n)^T$, and we can immediately calculate $X(X^TX)^{-1}X^T$ to obtain $h_{ij} = x_ix_j/\sum_{k=1}^{n} x_k^2$. Again $\sum_{i=1}^{n} h_i = p = 1$.

The usual regression line

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

has

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

and a few steps of algebra give

$$h_{ij} = \frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{\sum_{k=1}^{n}(x_k-\bar{x})^2} \quad . \tag{3.1}$$

Finally, we should 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 [4] discuss variances of residuals in several more complicated designs. It is straightforward to find H through equation (2.5).) The usual model for the $R \times C$ table is

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

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 $h_{ij,k\ell}$. 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

$$h_{ij,ij} = \frac{1}{C} + \frac{1}{R} - \frac{1}{RC} = \frac{R+C-1}{RC} \quad ; \qquad (3.3)$$

$$h_{ij,i\ell} = \frac{R-1}{RC} \quad , \quad \ell \neq j \quad ; \qquad (3.4)$$

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

$$h_{ij,k\ell} = -\frac{1}{RC} \quad , \quad k \neq i, \ \ell \neq j \quad . \qquad (3.6)$$

From equation (3.3) we see that all the diagonal elements of H are equal, as we would expect in a balanced design. It is worth mentioning, however, that such balance of leverage does not provide any particular "robustness" of fit. The appropriate notion is "resistance" -- a fit is resistant if a substantial change in only a small fraction of the data causes only a small change in the fit. Equations (3.3) through (3.6) show that two-way ANOVA is not resistant: $\hat{y}_{ij}$ will be affected by any change in $y_{k\ell}$ for any values of k and ℓ. If, instead of fitting by least squares, we were fitting by least absolute residuals (or by the related technique of median polish [10,12]), the result would be a resistant fit. This is true in part because the complete two-way table provides balance; the same degree of resistance is not in general to be found when fitting a simple straight line by least absolute residuals. Of course, such resistant alternatives

to least squares do not give rise to the hat matrix, which, together with other diagnostic tools, helps make possible effective data analysis by least squares. (We can expect, however, that future developments in resistant or robust linear fitting will yield their own analogues of H.) We turn now to two numerical examples showing the use of the hat matrix in multiple-regression situations.

4. NUMERICAL EXAMPLES

In this section we examine the hat matrix in two regression examples, emphasizing (either here or in Section 5) the connections between it and other sources of diagnostic information. We begin with a ten-point example, for which we can present H in full, and progress to a larger example, for which we shall work with only the diagonal elements, h_i .

The data for the first example comes from Draper and Stoneman [5]; we reproduce it in Exhibit 1. The response is strength, and the carriers are the constant, specific gravity, and moisture content. To probe the relationship between the non-constant carriers, we plot moisture content against specific gravity (Exhibit 2). 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

Exhibit 1

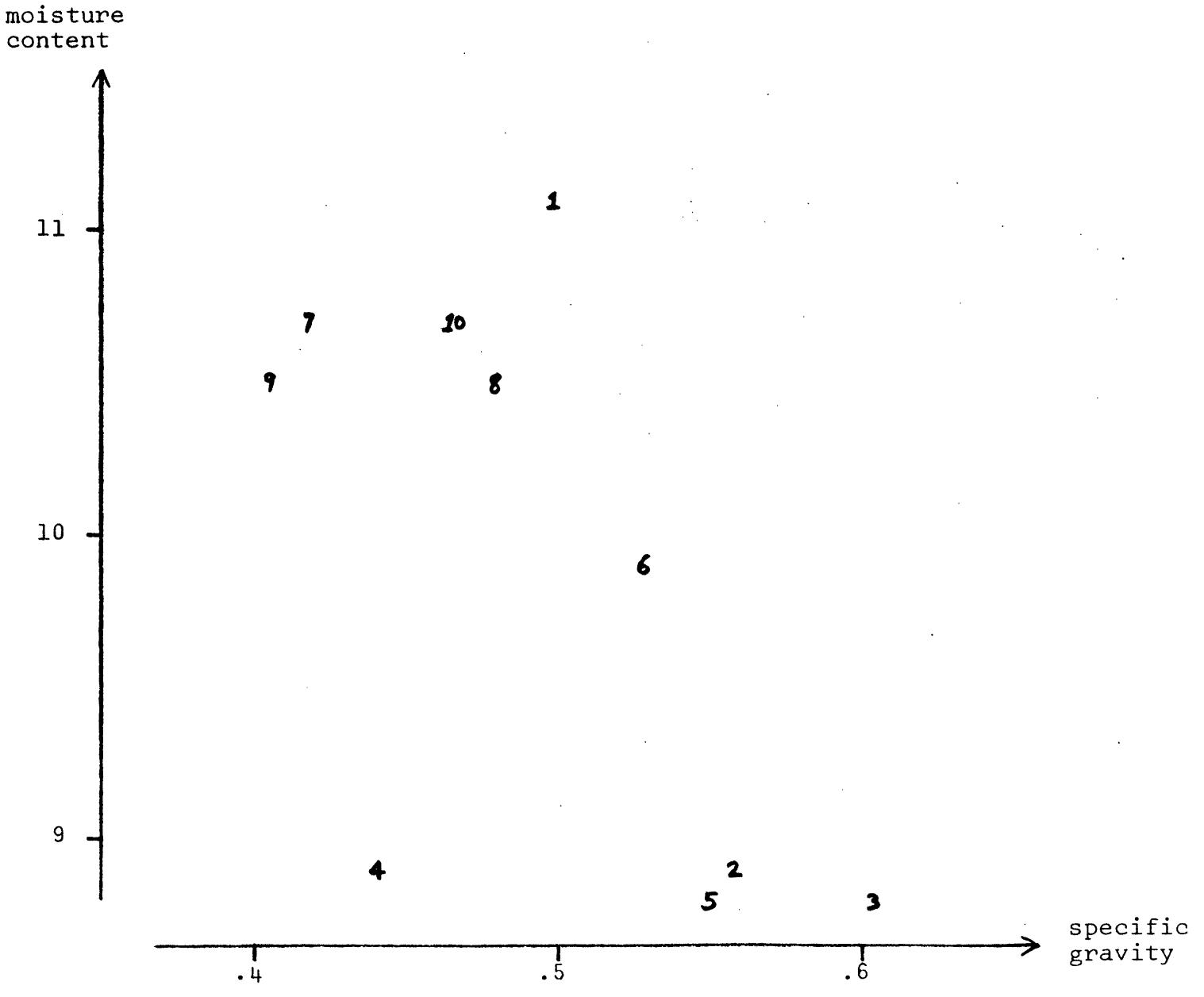
Data on Wood Beams

beam	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

Exhibit 2

The Two Carriers for the Wood Beam Data

(Plotting symbol is beam number.)



absent, it would be considerably more difficult to distinguish the two carriers.

The hat matrix for this X appears in Exhibit 3, and a stem-and-leaf display [11, 12] of the diagonal elements (rounded to multiples of .01) is as follows:

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

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 Exhibit 2. 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_{14} 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_{18} 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.

Exhibit 3

Hat Matrix for Wood Beam Data

(lower triangle omitted by symmetry)

i|j:	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

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 return to these questions in the next section. Now we turn to a larger example.

Our second example is based on savings rate data collected by Arlie Sterling of Massachusetts Institute of Technology. For purposes of illustration we use an econometric regression model for data of this type discussed by Leff [9]. Briefly, the life-cycle model of consumption implies that the aggregate propensity to save is related to the age distribution of the population, the level of real per capita disposable income, the rate of growth of real per capita disposable income, and other factors. For the 50 countries listed in Exhibit 4 the present set of data (Exhibit 5) consists of a response and four non-constant carriers and represents averages over the years 1960 through 1970. The response is a country's aggregate personal savings rate (abbreviated SR). The four carriers are the per cent of the population under age 15 (POP15), the per cent of the population over 75 (POP75), the level of real per capita disposable income measured in U.S. dollars (DILEV), and the per cent growth rate of DILEV (DIGRO).

In this example it is not too tedious to make and examine all pairwise scatter plots of the non-constant carriers. We include only two of the six scatter plots: Exhibit 6 shows DIGRO vs. POP15,

Exhibit 4

Country Labels for the Savings Rate Data

1	Australia	26	Malta
2	Austria	27	Norway
3	Belgium	28	Netherlands
4	Bolivia	29	New Zealand
5	Brazil	30	Nicaragua
6	Canada	31	Panama
7	Chile	32	Paraguay
8	China (Taiwan)	33	Peru
9	Colombia	34	Philippines
10	Costa Rica	35	Portugal
11	Denmark	36	South Africa
12	Ecuador	37	Southern Rhodesia
13	Finland	38	Spain
14	France	39	Sweden
15	Germany (F.R.)	40	Switzerland
16	Greece	41	Turkey
17	Guatemala	42	Tunisia
18	Honduras	43	United Kingdom
19	Iceland	44	United States
20	India	45	Venezuela
21	Ireland	46	Zambia
22	Italy	47	Jamaica
23	Japan	48	Uruguay
24	Korea	49	Libya
25	Luxembourg	50	Malaysia

Exhibit 5

Savings Rate Data

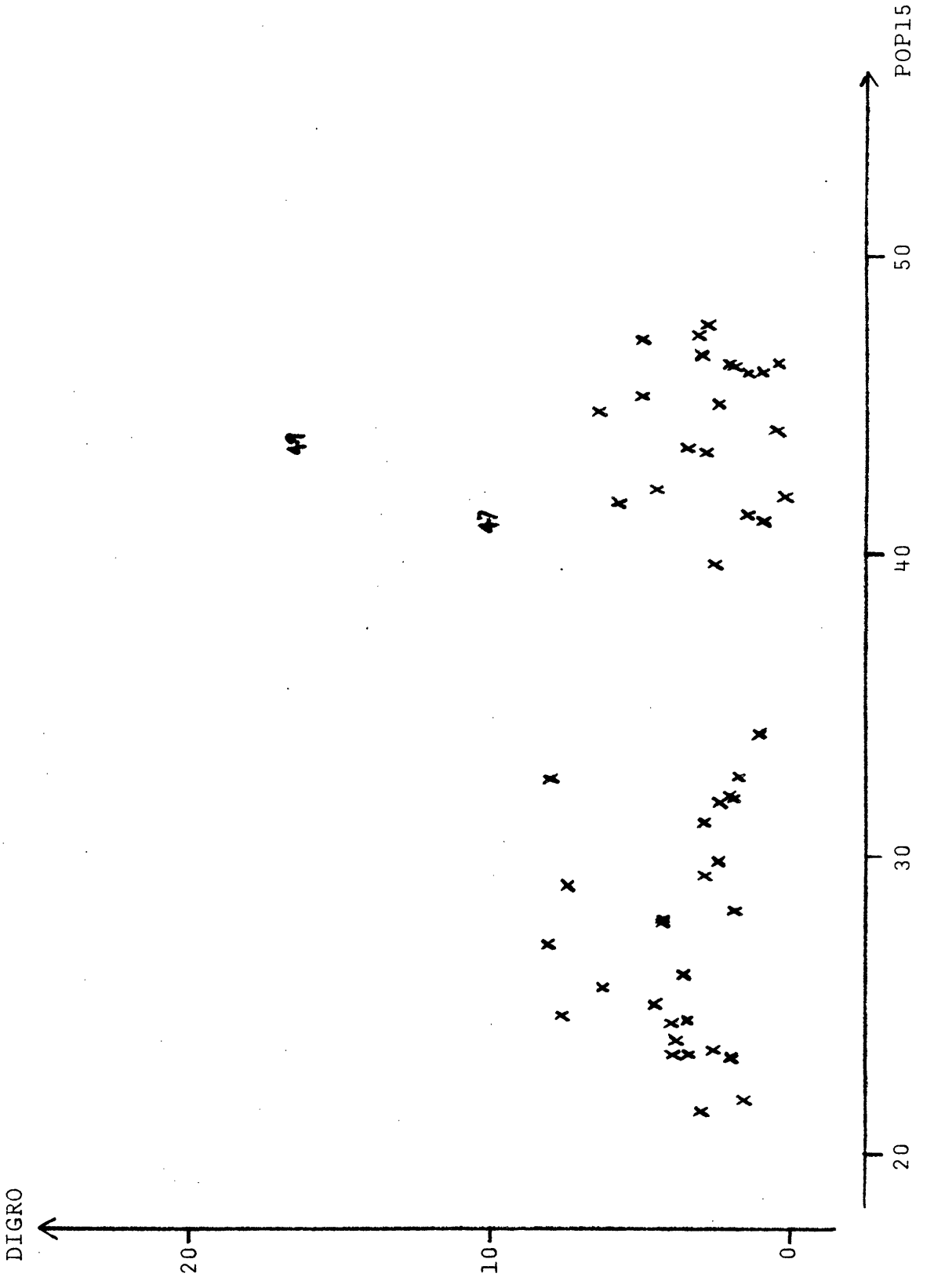
country	POP15	POP75	DILEV	DIGRO	SR
1	29.35	2.87	2329.68	2.87	11.43
2	23.32	4.41	1507.99	3.93	12.07
3	23.80	4.43	2108.47	3.82	13.17
4	41.89	1.67	189.13	0.22	5.75
5	42.19	0.83	728.47	4.56	12.88
6	31.72	2.85	2982.88	2.43	8.79
7	39.74	1.34	662.86	2.67	0.60
8	44.75	0.67	289.	6.51	11.90
9	46.64	1.06	276.	3.08	4.98
10	47.64	1.14	471.	2.80	10.78
11	24.42	3.93	2496.53	3.99	16.85
12	46.31	1.19	287.77	2.19	3.59
13	27.84	2.37	1681.25	4.32	11.24
14	25.06	4.70	2213.82	4.52	12.64
15	23.31	3.35	2457.12	3.44	12.55
16	25.62	3.10	870.85	6.28	10.67
17	46.05	0.87	289.71	1.48	3.01
18	47.32	0.58	232.44	3.19	7.70
19	34.03	3.08	1900.10	1.12	1.27
20	41.31	0.96	88.	1.54	9.00
21	31.16	4.19	1139.	2.99	11.34
22	24.52	3.48	1390.00	3.54	14.28
23	27.01	1.91	1257.28	8.21	21.10
24	41.74	0.91	207.68	5.81	3.98
25	21.80	3.73	2449.39	1.57	10.35

Exhibit 5 continued

country	POP15	POP75	DILEV	DIGRO	SR
26	32.54	2.47	601.05	8.12	15.48
27	25.95	3.67	2231.03	3.62	10.25
28	24.71	3.25	1740.70	7.66	14.65
29	32.61	3.17	1487.52	1.76	10.67
30	45.04	1.21	325.54	2.48	7.30
31	43.56	1.20	568.56	3.61	4.44
32	41.18	1.05	220.56	1.03	2.02
33	44.19	1.28	400.06	0.67	12.70
34	46.26	1.12	152.01	2.00	12.78
35	28.96	2.85	579.51	7.48	12.49
36	31.94	2.28	651.11	2.19	11.14
37	31.92	1.52	250.96	2.00	13.30
38	27.74	2.87	768.79	4.35	11.77
39	21.44	4.54	3299.49	3.01	6.86
40	23.49	3.73	2630.96	2.70	14.13
41	43.42	1.08	389.66	2.96	5.13
42	46.12	1.21	249.87	1.13	2.81
43	23.27	4.46	1813.93	2.01	7.81
44	29.81	3.43	4001.89	2.45	7.56
45	46.40	0.90	813.39	0.53	9.22
46	45.25	0.56	138.33	5.14	18.56
47	41.12	1.73	380.47	10.23	7.72
48	28.13	2.72	766.54	1.88	9.24
49	43.69	2.07	123.58	16.71	8.89
50	47.20	0.66	242.69	5.08	4.71

Exhibit 6

Savings Rate Data: DIGRO vs. POP15



while Exhibit 7 plots DILEV against POP75. Examination of all six plots led us to regard Libya (country 49) and perhaps Jamaica (47) and the United States (44) as unusual.

Exhibit 8 gives a stem-and-leaf display of the 50 diagonal elements of the hat matrix (whose full numerical values appear in Exhibit 14). Since $2p/n = .2$ here, we identify Ireland (21), Japan (23), the United States (44), and Libya (49) as high-leverage points. We investigate their influence on the estimated coefficients and on the fitted values in the next section.

5. BRINGING IN THE RESIDUALS

So far we have examined the design matrix X for evidence of points with 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 (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. Alternatively, the accuracy of the apparently discrepant point may

Exhibit 7

Savings Rate Data: DILEV vs. POP75

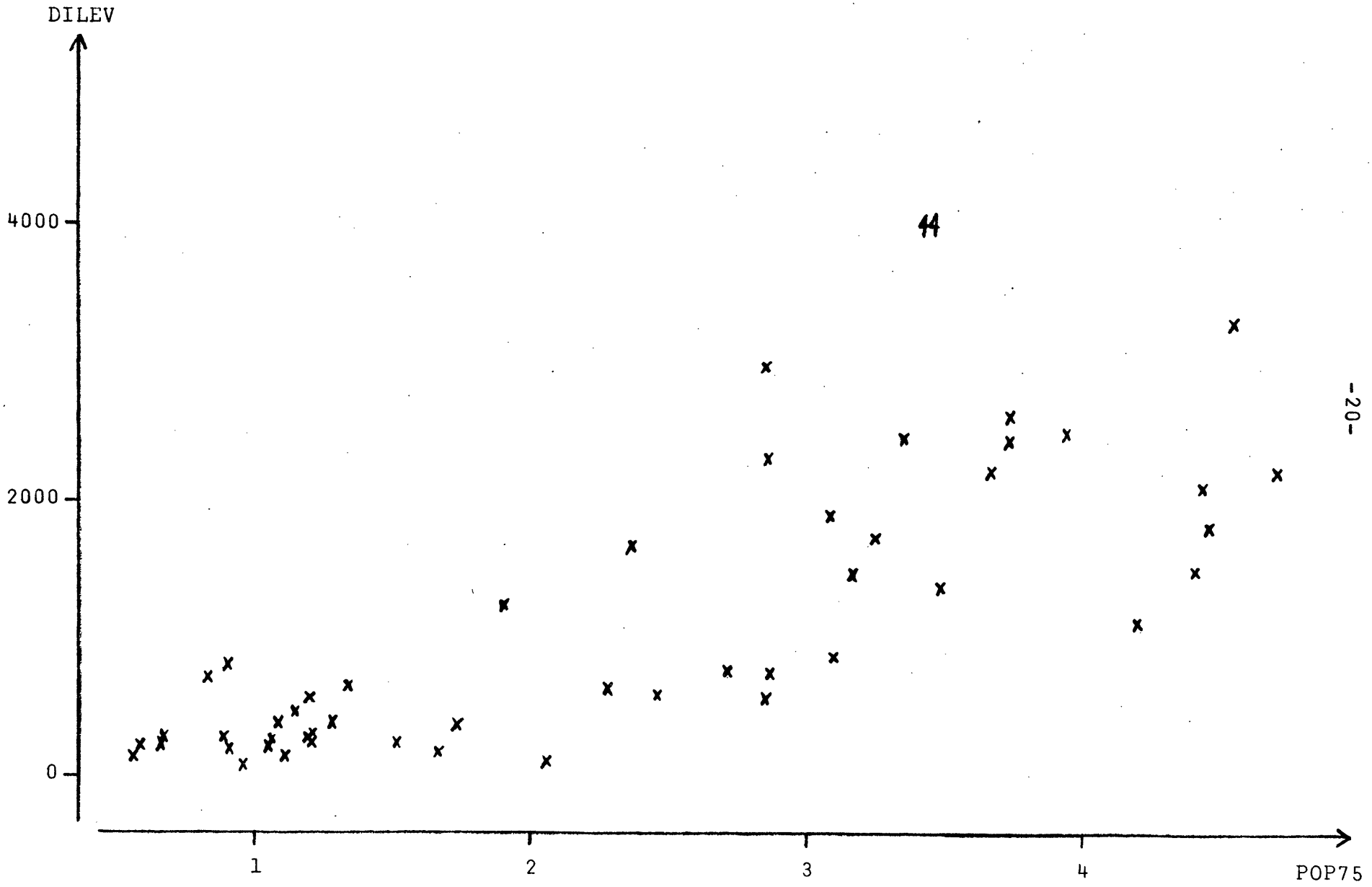


Exhibit 8

Diagonal Elements of H for Savings Rate Data

stem-and-leaf display

(unit = .001)

3	789
4	7
5	047
6	00023445556799
7	01345779
8	66779
9	02677
10	
11	6
12	03
13	6
14	0
15	8
16	0
hi	.212, .223, .333, .531

	↑	↑	↑	↑
country tags:	21	23	44	49

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 has discussed and illustrated some of these [1].) When there is substantial variation among the h_i values, equation (2.5) indicates that we should allow for differences in the variances of the r_i [2] and look at $r_i/\sqrt{1-h_i}$. This adjustment puts the residuals on an equal footing, but it is often more convenient to use the <u>standardized residual</u>, $r_i/(s\sqrt{1-h_i})$, 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. Denoting row i of X, that is, $(x_{i1},\ldots,x_{ip})$, by $\underset{\sim}{x}_i$, this residual is $y_i - \underset{\sim}{x}_i\hat{\underset{\sim}{\beta}}_{(i)}$, where $\hat{\underset{\sim}{\beta}}_{(i)}$ is the least-squares estimate of $\underset{\sim}{\beta}$ based on all the data except data point i. Similarly $s^2_{(i)}$ is the residual mean square for the "not-i" fit, and the standard deviation of $y_i - \underset{\sim}{x}_i\hat{\underset{\sim}{\beta}}_{(i)}$ is estimated by $s_{(i)}\sqrt{1 + \underset{\sim}{x}_i(X^T_{(i)}X_{(i)})^{-1}\underset{\sim}{x}_i^T}$. ($X_{(i)}$ is obtained from X by deleting row i.) We now define the <u>studentized residual</u>:

$$r_i^* = \frac{y_i - \underset{\sim}{x}_i \hat{\underset{\sim}{\beta}}_{(i)}}{s_{(i)}\sqrt{1 + \underset{\sim}{x}_i (X_{(i)}^T X_{(i)})^{-1} \underset{\sim}{x}_i^T}} \quad . \qquad (5.1)$$

Since the numerator and denominator in (5.1) 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 turns (5.1) into

$$r_i^* = r_i/(s_{(i)}\sqrt{1-h_i}) \qquad (5.2)$$

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

$$(n-p-1)s_{(i)}^2 = (n-p)s^2 - \frac{r_i^2}{1-h_i} \quad . \qquad (5.3)$$

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. When h_i is large, it may still be the case that r_i^* is moderate or small (because y_i is not discrepant or because it has exerted its leverage on the fit), and we must determine the impact of such points by setting them aside and refitting without them.

Thus by examining the h_i we are able to find troublesome points which we might miss if we used only the studentized residuals. Since we have already discussed the h_i for our two numerical examples, we now turn to their studentized residuals.

For the wood beam example, we plot strength against specific gravity in Exhibit 9 and strength against moisture content in Exhibit 10. 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 (which we have already flagged as high-leverage) stands apart from the rest. Exhibit 11 gives r_i , $\sqrt{1-h_i}$, $s_{(i)}$, and the studentized residuals r_i^* . Among the r_i^* , beam 1 appears as a clear stray ($p<.02$), and beam 6 may also deserve attention ($p<.1$). Since beam 4 is known to have high leverage ($h_i=.604$), we should still be suspicious of it, even though r_4^* is not particularly large. The fit for the full data is

$$\hat{y} = 10.302 + 8.495(SG) - 0.2663(MC) \qquad (5.4)$$

with $s = 0.275$; and when we set aside beam 4, the fit changes to

$$\hat{y} = 12.411 + 6.799(SG) - 0.3905(MC) \quad , \qquad (5.5)$$

a noticeable shift. To judge the importance of these coefficient changes, we must consider the variability of the estimates. The most convenient source for this information is the covariance matrix of $\hat{\underset{\sim}{\beta}}$, which is equal to $s^2(X^TX)^{-1}$. In this case $s^2 = 0.07578$, and

Exhibit 9

Wood Beam Data -- Strength vs. Specific Gravity

(Plotting symbol is beam number.)

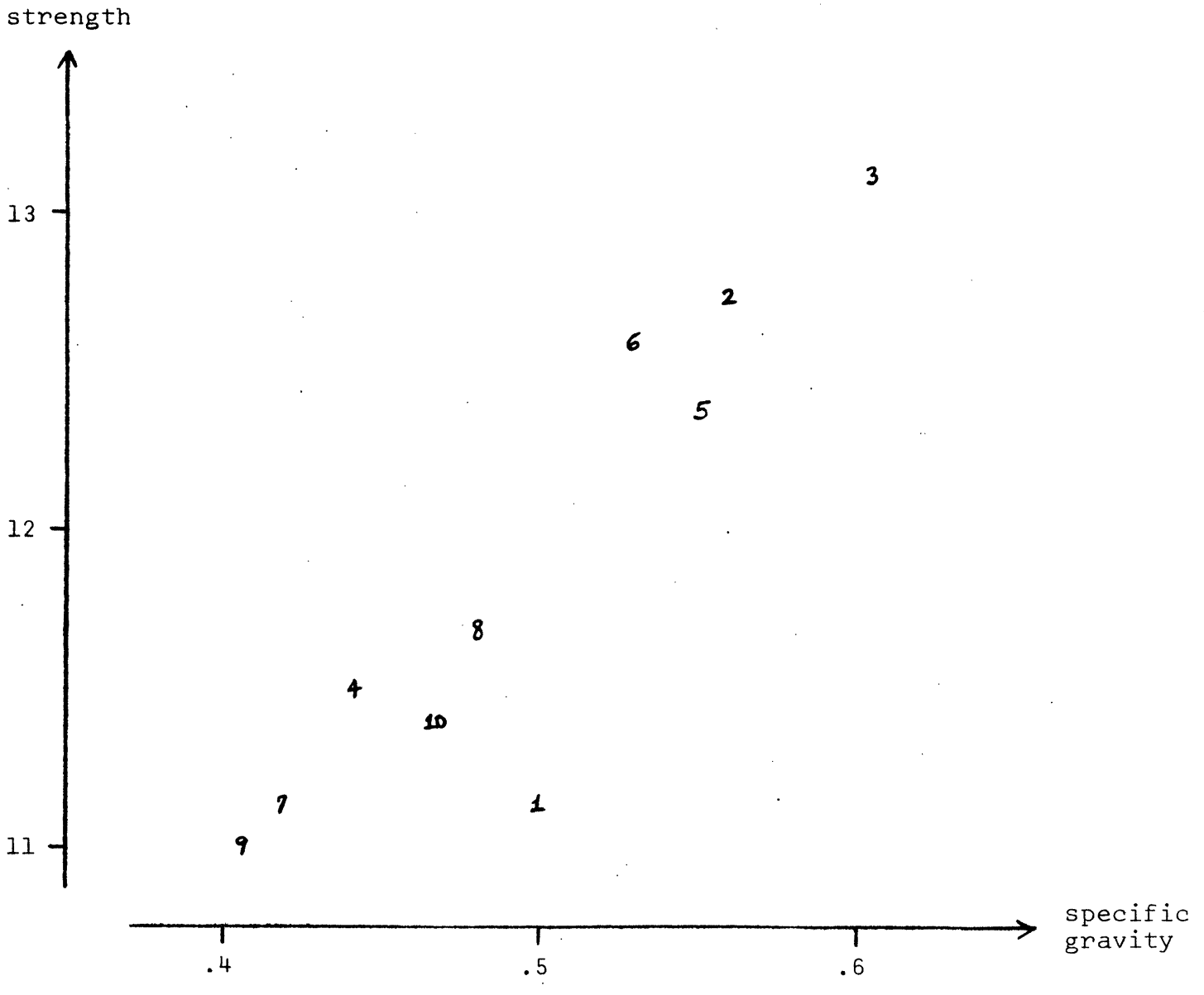


Exhibit 10

Wood Beam Data -- Strength vs. Moisture Content

(Plotting symbol is beam number.)

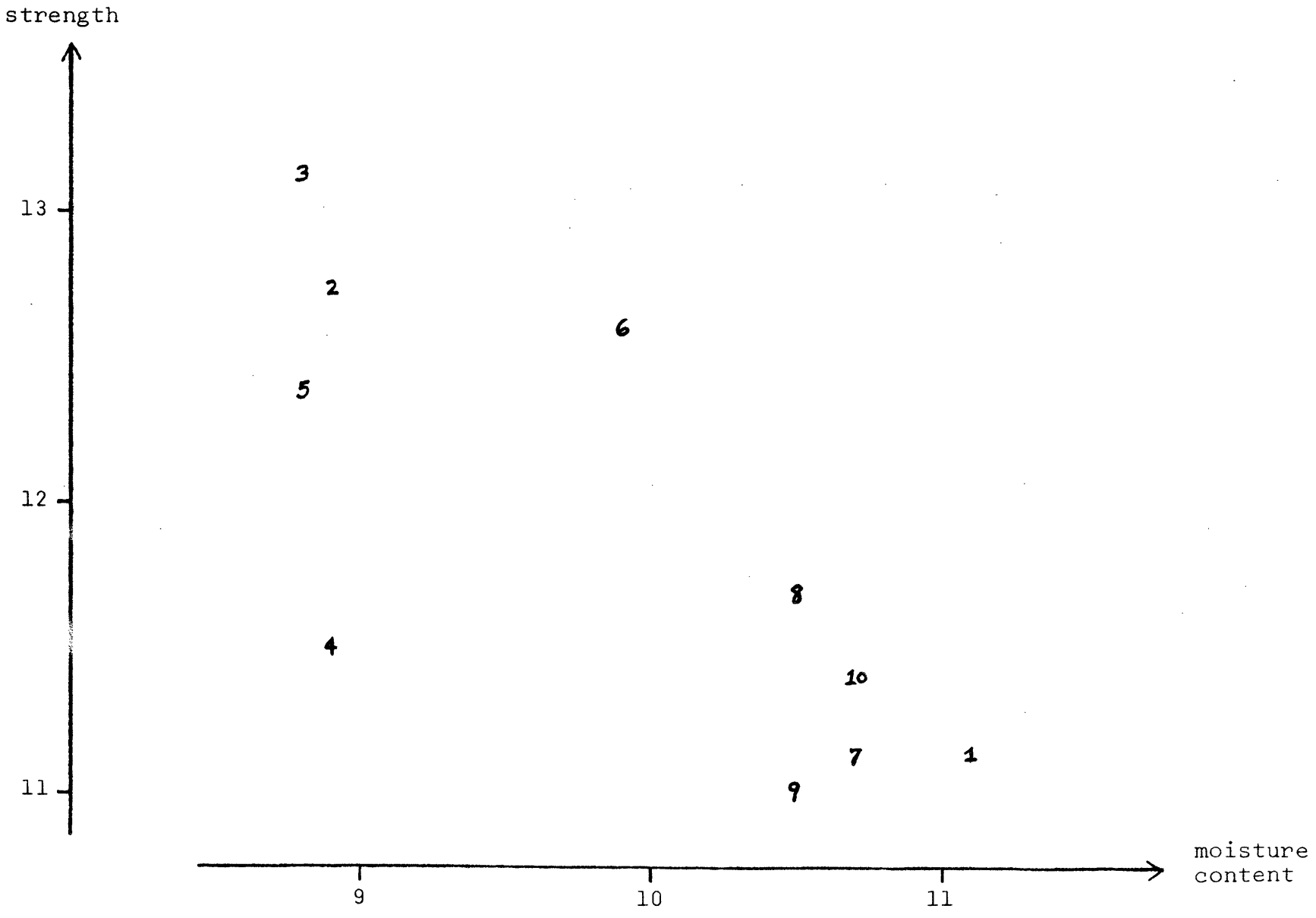


Exhibit 11

Studentized Residuals and Related Quantities

(wood beam data)

i	r_i	h_i	$\sqrt{1-h_i}$	$s_{(i)}$	r_i^*
1	-.448	.418	.763	.176	-3.338
2	.065	.242	.871	.296	.252
3	.038	.417	.764	.297	.168
4	-.171	.604	.629	.276	- .985
5	-.253	.252	.865	.272	-1.074
6	.446	.148	.923	.222	2.172
7	.123	.262	.859	.292	.491
8	.113	.154	.920	.293	.419
9	.062	.315	.828	.296	.253
10	-.013	.187	.902	.297	.048

$$(X^TX)^{-1} = \begin{bmatrix} 47.408 & -38.275 & -2.870 \\ -38.275 & 41.998 & 1.769 \\ -2.870 & 1.769 & 0.202 \end{bmatrix} .$$

Thus the coefficient changes from (5.4) to (5.5) are, in standard-error units, 0.306, -0.262, and -0.276, respectively. Whether we take these individually or as a whole, we are not led to conclude that beam 4 is seriously discrepant. We could examine the effect of setting aside beam 1 and possibly beam 6, but we do not pursue this here.

For the savings rate data we can examine the plots of the response against each carrier as in the wood beam example. In one of these, Exhibit 12, we plot SR against POP15 and see that Zambia (46) and Japan (23) are notable but that Lybia (49), a point we have also flagged, is not. The same three points are marked in Exhibit 13, which plots SR against DIGRO. Point 49 is again notable, but it is hard to say from this plot alone how much it affects the multiple regression fit.

Turning next to the studentized residuals r_i^* (in Exhibit 14), we can use the value 2 (approximately the two-sided 95% point of t_{44}) as a rough cut-off, finding Chile (7) and Zambia (46) discrepant. In this example, analysis of the residuals does not reveal any of the high-leverage points.

To assess the impact of these four leverage points, we compute the change in the coefficients when each of the points is removed. The formula

Exhibit 12

Savings Rate Data: SR vs. POP15

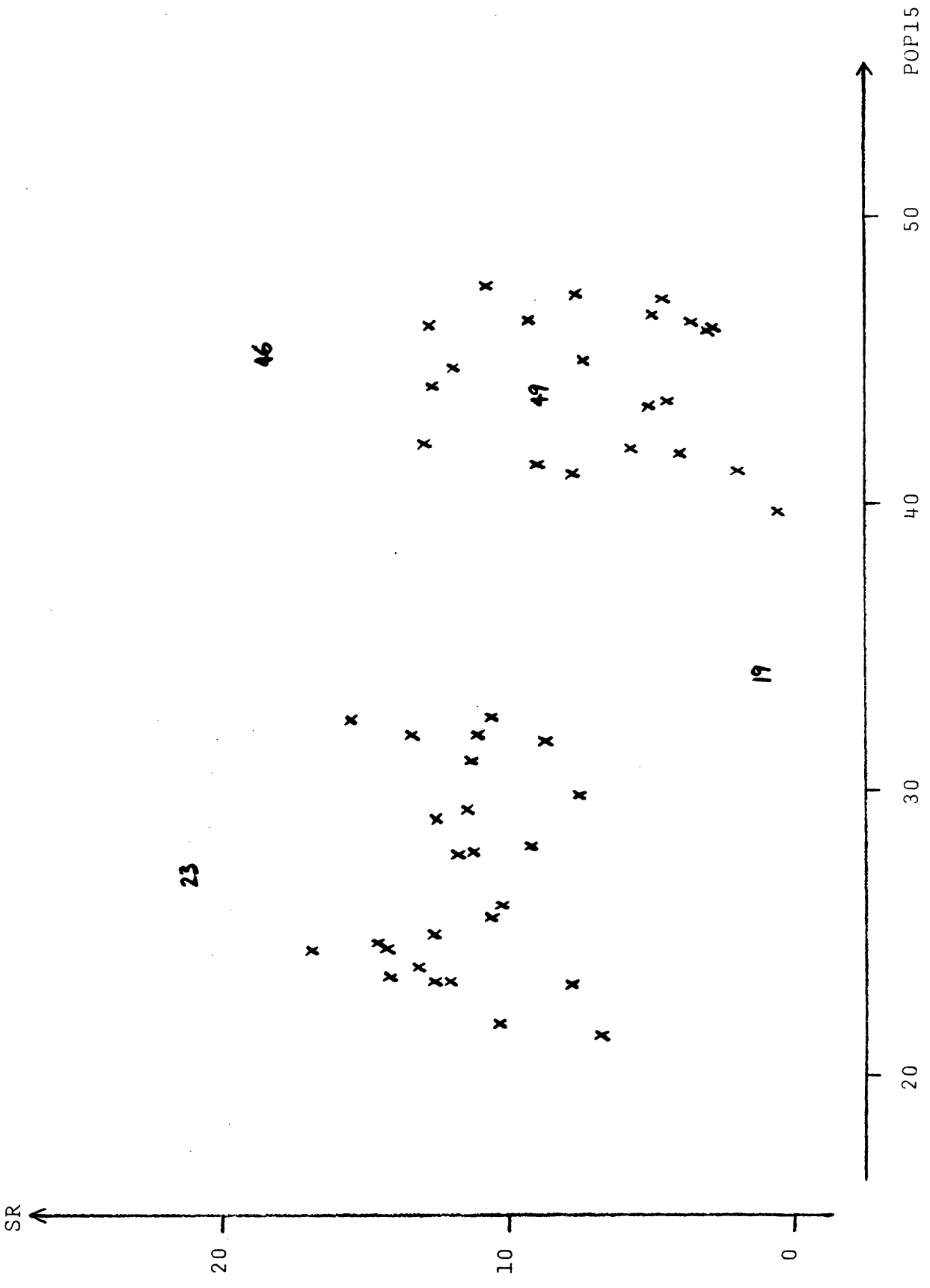


Exhibit 13

Savings Rate Data: SR vs. DIGRO

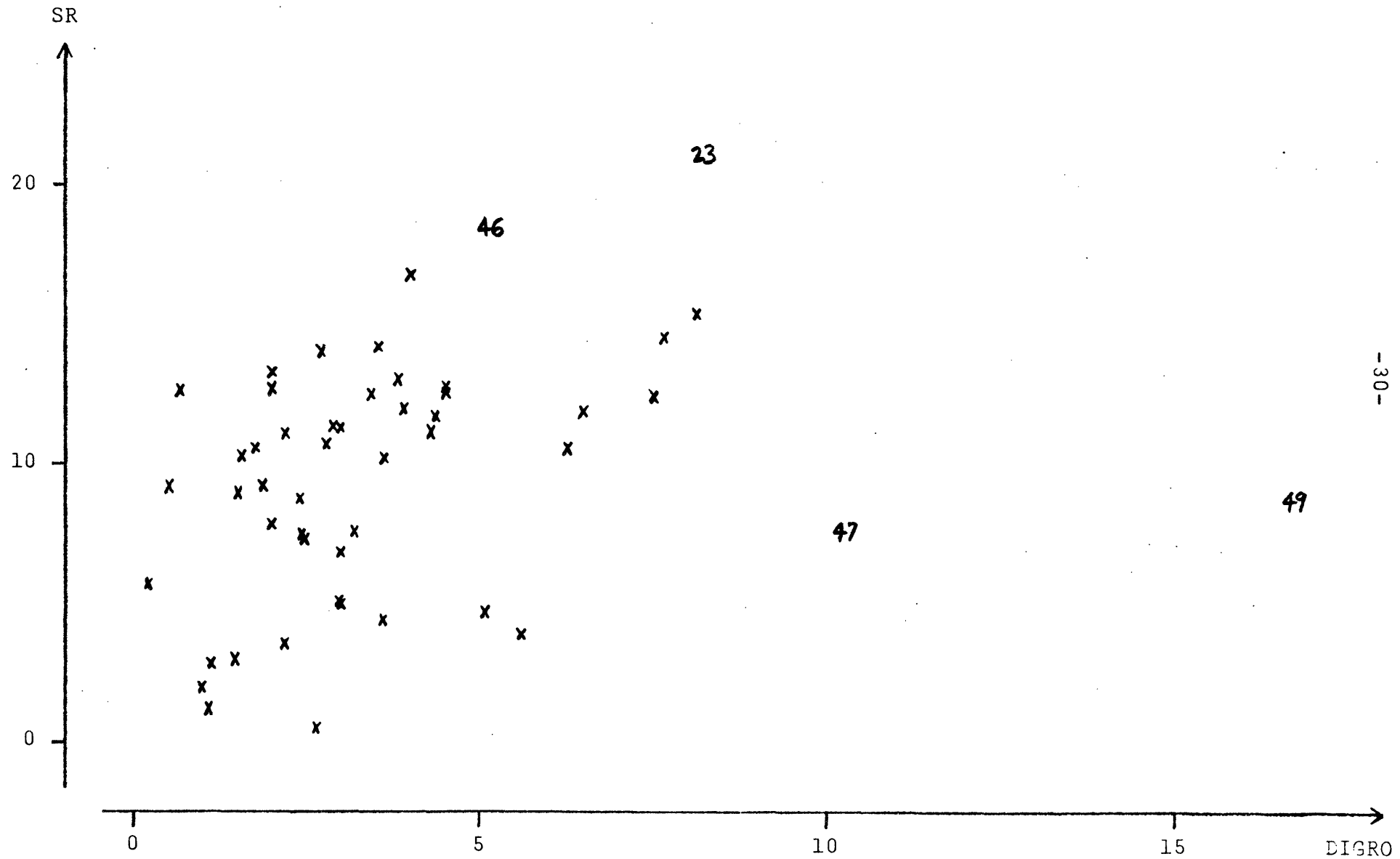


Exhibit 14

Studentized Residuals for the Savings Rate Data

i	r_i	h_i	$\sqrt{1-h_i}$	$s_{(i)}$	r_i^*
1	0.864	.0677	.9656	3.843	0.233
2	0.616	.1204	.9379	3.844	0.171
3	2.219	.0875	.9553	3.830	0.607
4	-0.698	.0895	.9542	3.844	-0.190
5	3.553	.0696	.9646	3.805	0.968
6	-0.317	.1584	.9174	3.845	-0.090
7	-8.242	.0373	.9812	3.631	-2.313
8	2.536	.0780	.9602	3.825	0.690
9	-1.452	.0573	.9709	3.839	-0.389
10	5.125	.0755	.9615	3.761	1.417
11	5.400	.0627	.9681	3.753	1.486
12	-2.406	.0637	.9676	3.827	-0.650
13	-1.681	.0920	.9529	3.836	-0.460
14	2.475	.1362	.9294	3.825	0.696
15	-0.181	.0874	.9553	3.846	-0.049
16	-3.116	.0966	.9505	3.814	-0.860
17	-3.355	.0605	.9693	3.810	-0.909
18	0.710	.0601	.9695	3.844	0.191
19	-6.211	.0705	.9641	3.721	-1.731
20	0.509	.0715	.9636	3.845	0.137
21	3.391	.2122	.8876	3.802	1.005
22	1.927	.0665	.9662	3.834	0.520
23	5.281	.2233	.8813	3.738	1.603
24	-6.107	.0608	.9691	3.726	-1.691
25	-1.671	.0863	.9559	3.837	-0.456

Exhibit 14 continued

i	r_i	h_i	$\sqrt{1-h_i}$	$s_{(i)}$	r_i^*
26	2.975	.0794	.9595	3.817	0.812
27	-0.872	.0479	.9757	3.843	-0.232
28	0.426	.0906	.9536	3.845	0.116
29	2.286	.0542	.9725	3.829	0.614
30	0.646	.0504	.9745	3.844	0.173
31	-3.294	.0390	.9803	3.812	-0.881
32	-6.126	.0694	.9647	3.725	-1.705
33	6.539	.0650	.9669	3.708	1.824
34	6.675	.0643	.9673	3.702	1.864
35	-0.768	.0971	.9502	3.844	-0.210
36	0.483	.0651	.9669	3.845	0.130
37	1.291	.1608	.9161	3.840	0.367
38	-0.671	.0773	.9606	3.844	-0.182
39	-4.260	.1240	.9360	3.784	-1.203
40	2.487	.0736	.9625	3.826	0.675
41	-2.666	.0396	.9800	3.824	-0.711
42	-2.818	.0746	.9620	3.820	-0.767
43	-2.692	.1165	.9399	3.821	-0.750
44	-1.112	.3337	.8163	3.840	-0.355
45	3.633	.0863	.9559	3.803	0.999
46	9.751	.0643	.9673	3.533	2.854
47	-3.019	.1408	.9270	3.814	-0.854
48	-2.264	.0979	.9498	3.829	-0.623
49	-2.830	.5315	.6845	3.795	-1.089
50	-2.971	.0652	.9668	3.818	-0.805

Note: For country labels, see Exhibit 4.

$$\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_{(i)} = (X^T X)^{-1} \underset{\sim}{x}_i^T \, r_i/(1-h_i) \qquad (5.6)$$

simplifies this considerably. Exhibit 15 gives the changes for countries 49, 44, 23, and 21 along with the components of $\hat{\underset{\sim}{\beta}}$ and their standard errors. Since removal of Libya (49) causes the coefficient of DIGRO to change by more than one standard error, we should be cautious about including that data point. In contrast, removing the United States (44) has little impact on $\hat{\underset{\sim}{\beta}}$, and thus this country appears to be consistent with the rest of the data. Such a leverage point should usually be retained because it can play an important role in limiting the variances and covariances of coefficient estimates. The other two high-leverage countries, Japan (23) and Ireland (21), do not appear to be especially influential, but we have lost very little by checking to make sure.

In both examples 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

Exhibit 15

Coefficient Changes When Individual

Data Points Are Omitted

(Entries are components of $\hat{\beta}-\hat{\beta}_{(i)}$.)

country omitted	carrier				
	CONST	POP15	POP75	DILEV	DIGRO
49	4.042	.0698	-.4106	-.000018	-.2005
44	0.513	-.0106	.0410	-.000219	-.0065
23	4.626	-.0933	-.7178	.000134	.0749
21	-2.280	.0428	.5218	-.000240	-.0183
$\hat{\beta}$	28.566	-.4612	-1.6915	.000337	.4097
s.e.	7.354	.1446	1.0836	.000931	.1962

Note: CONST is the constant carrier, whose value is always 1 .

more insightful analysis than if we had followed a routine or automated pattern of 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 [5] 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" [10].

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

$$\underset{n\times p}{X} = \underset{n\times n}{Q}\ \underset{n\times p}{R} \qquad (6.1)$$

(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

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

With a modest increase in computation time and/or storage, a simple modification of the basic algorithm yields H as a by-product. If n is large, we can use a somewhat different modification to calculate and store only the h_i.

Some least-squares solvers use the modified Gram-Schmidt algorithm to find a different QR-factorization of X:

$$\underset{n \times p}{X} = \underset{n \times p}{Q} \quad \underset{p \times p}{R} \quad . \tag{6.3}$$

Here $Q^TQ = I_p$ and R is upper triangular, and it is easy to see that

$$H = QQ^T \quad . \tag{6.4}$$

It is possible to build up the h_i during the calculation without storing Q, but modified Gram-Schmidt is not as accurate for this as it is for determining the least-squares estimate of $\hat{\beta}$.

Finally we mention the singular-value decomposition,

$$\underset{n\times p}{X} = \underset{n\times p}{U}\ \underset{p\times p}{\sum}\ \underset{p\times p}{V^T} \quad , \tag{6.5}$$

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

$$H = UU^T \quad . \tag{6.6}$$

These and other decompositions are discussed in [6]. For a recent account of numerical techniques in solving linear least-squares problems, we recommend the book by Lawson and Hanson [8].

Appendix

In this appendix we formally show that when $h_1=1$ (we can take $i=1$ without loss of generality), there exists a nonsingular transformation T, such that $\hat{\alpha}_1 = (T^{-1}\hat{\underset{\sim}{\beta}})_1 = y_1$ and $\hat{\alpha}_2,\ldots,\hat{\alpha}_p$ do not depend on y_1. This implies that, in the transformed coordinate system, the parameter α_1 has been dedicated to observation 1.

When $h_1=1$, we have for the coordinate vector $\underset{\sim}{e}_1 = (1,0,\ldots,0)^T$

$$H\underset{\sim}{e}_1 = \underset{\sim}{e}_1$$

since (2.6) shows that $h_{1j} = 0$, $j\neq 1$. Let P be any $p\times p$ nonsingular matrix whose first column is $(X^TX)^{-1}X^T\underset{\sim}{e}_1$. Then

$$XP = \begin{bmatrix} 1 & \underset{\sim}{a} \\ \underset{\sim}{0} & A \end{bmatrix}$$

where $\underset{\sim}{a}$ is $1\times(p-1)$ and $\underset{\sim}{0}$ is $(p-1)\times 1$. Now let

$$Q = \begin{bmatrix} 1 & -\underset{\sim}{a} \\ \underset{\sim}{0} & I \end{bmatrix}$$

with I denoting the $(p-1)\times(p-1)$ identity matrix. The transformation we seek is given by $T = PQ$, which is nonsingular because both P and Q have inverses. Clearly

$$XT = \begin{bmatrix} 1 & \underset{\sim}{0} \\ \underset{\sim}{0} & A \end{bmatrix} ,$$

and the least-squares estimate of the parameter $\underset{\sim}{\alpha} = T^{-1}\underset{\sim}{\beta}$ will have the first residual, $y_1-\hat{\alpha}_1$, equal to zero since $\hat{\alpha}_2,\ldots,\hat{\alpha}_p$ cannot affect this residual. This also implies that $\hat{\alpha}_2,\ldots,\hat{\alpha}_p$ will not depend on y_1 .

REFERENCES

[1] Anscombe, F.J. (1973). "Graphs in Statistical Analysis," The American Statistician, 27, 17-21.

[2] Anscombe, F.J. and Tukey, J.W. (1963). "The Examination and Analysis of Residuals," Technometrics, 5, 141-160.

[3] Beckman, R.J. and Trussell, H.J. (1974). "The Distribution of an Arbitrary Studentized Residual and the Effects of Updating in Multiple Regression," Journal of the American Statistical Association, 69, 199-201.

[4] Behnken, D.W. and Draper, N.R. (1972). "Residuals and Their Variance Patterns," Technometrics, 14, 101-111.

[5] 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.

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

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

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

[9] Leff, N.H. (1969). "Dependency Rates and Savings Rates," American Economic Review, 59, 886-896.

[10] Tukey, J.W. (1972). "Data Analysis, Computation, and Mathematics," Quarterly of Applied Mathematics, 30, 51-65.

[11] Tukey, J.W. (1972). "Some Graphic and Semigraphic Displays," in Statistical Papers in Honor of George W. Snedecor (ed. T.A. Bancroft), Ames, Iowa: Iowa State University Press.

[12] Tukey, J.W. (1977). Exploratory Data Analysis, Reading, Mass: Addison-Wesley.