ACTA ASTRONOMICA
Vol. 39 (1989) pp. 257-273

Investigation of the Influential Points in a Regression Problem Occurring in Short-Term Prediction of Solar Flare Activity

by

M. Jakimiec

Astronomical Institute, University of Wrocław, Poland

and

A. Bartkowiak

Institute of Computer Science, University of Wrocław, Poland

Received December 1, 1988

ABSTRACT

We consider the regression equations employed often for short-term flare activity predictions. In these regressions the sunspot group characteristics on the given day and the flare activity characteristics on the next day are taken into account. The regression functions are estimated from the empirical data, in which some atypical data vectors may be contained. The problem we dealt in this paper is: How much the regression functions used as predicting algorithms are influenced by such single, atypical data vector. To solve the problem the stability of the regressions are analysed. Four regression diagnostics are used to make clear the impact of singular data vectors on the regression equations. These diagnostics allow us to identify atypical data vectors which characterize sunspot groups e.g. instantaneously changing flare activity level or such ones with extremaly high flare activity. We have found that the predicting functions are sufficiently stable, i.e. that for the considered empirical data they are not disturbed strongly by such singular data vectors.

1. Introduction

When considering a regression equation in a multivariate problem one should be sure that its coefficients are stable. Otherwise, the result of the regression analysis could be put in doubt. The coefficients of a regression equation are stable if they do not depend too strongly on the particular data set sampled from the analysed population. In the problem of short-term prediction of solar flare activity many investigators have applied different methods, mainly regression analysis methods. Some of the results (e.g. Hirman *et al.* 1980, Jakimiec and Wasiucionek 1980, Neidig *et al.* 1986) indicate that the estimated predicting function coefficients are unstable. The instability of the regression function can arise from various sources. For instance, the impact of atypical data vectors included into the data set on the predicting function coefficients might be such a source. This kind of the instability problem, accordingly to our knowledge, has not been examinated as yet.

Some methods allowing to reveal atypical data vectors are presented by Bartkowiak and Jakimiec (1989). They have found that a big error in one of the data vectors can change quite considerably the structure of the covariances between the considered variables. In the present paper we apply some statistical methods allowing to investigate the stability of a regression equation. In the following we will consider four statistics (statistical indices called also regression diagnostics) showing the impact of single data vector on the stability of the regression equation. These are: 1. The diagonal elements of the HAT matrix h; 2. Externally studentized residuals t; 3. A statistics introduced by Belsley *et al.* (1980), called by them DFFITS; 4. Mean percent of distortion of the regression coefficients, called by us MDB. In all presented here methods we use the leaving-one-out technique, i.e. we consider some statistics evaluated consecutively n times after removing each time one data vector (individual) from the data.

2. The data

Generally, in our investigations a data vector (an item) comprises p explanatory variables $X_1, X_2, \ldots, X_p$, characterizing a sunspot group on a given day and one predicting variable Y characterizing the flare activity of the given sunspot group on the next day. We analyse in this work essentialy the same data as used by Bartkowiak and Jakimiec (1989), i.e. we analyse a complex of thirteen explanatory variables $X_1, \ldots, X_{13}$, characterizing sunspot groups of D, E, F Zurich classes and being in the decay phase of the evolution. We omit now the variable $x14$ because it is strongly correlated with the variable $x12$ and does not add any new information. However, in this work we consider two additional predicted variables $Y(Fs$ and $Fh)$ describing the sunspot group flare activity on the next day: the daily sums of the X-ray flare fluxes in the wavelength intervals $1-8$ Å (Fs) and $0.5-4$ Å (Fh). Similarly as for the appropriate variables $x11$ and $x13$ we use here the logarithmic transformations $Y' = \log Y$ and $Y' = \log Y+2$ for the characteristics Fs and Fh.

We consider the regression:

$$y = b_0 + \sum_{j=1}^{p} b_j x_j + e, \tag{1}$$

with $p = 13$, the number of the explanatory or predicting variables, and taking in turn Fs and Fh as the predicted or explained variable Y, appropriately.

We have three sets of data. The first (set E) comprises erroneous values (due to errors in punching the data, as discussed in Bartkowiak and Jakimiec, 1989).

The second (set C) comprises the data corrected for the punching errors. The third (set A) was obtained after removing from the set C two atypical data vectors. Comparing the results obtained from the sets E and C we want to study the impact of errors on the regression. Comparing the results obtained from the data sets C and A we want to study especially the impact of two atypical data vectors on the regression.

3. Statistical methods

3.1. The diagonal elements of the HAT matrix as indicators of leverage points for a regression

The regression equation (1) can be presented in more general form:

$$y = Xb + e, \qquad (2)$$

where $y = (y_1, \ldots, y_n)$' is the vector of the values of the predicted variable; n is the number of considered data vectors (size of the data set); $X = (x_{ij})$ $(i = 1, \ldots, n; j = 0, \ldots, p)$ is the "design matrix" comprising in the first column a vector of "ones", and in the remaining columns the values of the predicting variables; p is the number of predicting variables; $b = (b_0, b_1, \ldots, b_p)$' is the vector of the regression coefficients; $e = (e_1, \ldots, e_n)$' is the vector of errors or inadequacies of the fit of the assumed model given by (1) or (2). Generally, it is assumed that the components of e are independent and identically distributed with mean values $E(e_i) = 0$ and variances $Var(e_i) = \sigma^2$ (for $i = 1, \ldots, n$). Further, it will be assumed that each e_i is normally distributed, i.e. $e_i \sim N(0, \sigma^2)$.

The least squares estimator $\hat{b}$ of the regression coefficients b appearing in (2) can be calculated by the formula:

$$\hat{b} = (X'X)^{-1}X'y, \qquad (3)$$

(provided the matrix X is of rank $m = p+1$, and $n > m$), wherefrom $\hat{y}$, the vector of expected values, can be obtained as:

$$\hat{y} = X\hat{b} = X(X'X)^{-1}X'y.$$

The HAT matrix is defined as $H = (h_{ij})(i,j = 1, 2, \ldots, n)$ obtained from the design matrix X by the formula:

$$H = X(X'X)^{-1}X'. \qquad (4)$$

It permits to obtain from the values of the predicted variable y the values of $\hat{y}$ expected from the assumed regression equation by using simple formulae:

$$\hat{y} = Hy. \qquad (5)$$

For example, assuming $i = 1$, i.e. considering the first individual (data vector) we obtain from (5) the following formula for computing the expected value $\hat{y}_1$:

$$\hat{y}_1 = h_{11}y_1 + h_{12}y_2 + \ldots + h_{1n}y_n.$$

The value h_{11} can be here interpreted as the amount of leverage or influence excerted by the value y_1 on $\hat{y}_1$. In general, the matrix H reveals points of high influence of position in the design. Usually these are values of $X_1, \ldots, X_p$ located at extreme positions of the cloud of the data points, which can influence strongly the regression equation and therefore special attention should be paid to such points. However, it can happen that points with high leverage do not influence the regression equation at all. For a discussion on this topic see Chatterjee and Hadi (1986).

It can be shown that the elements h_{ij} satisfy the inequalities:

$$0 \leqslant h_{ij} \leqslant 1, \tag{6}$$

since H is an idempotent matrix satisfying $H^2 = H$. Moreover, the trace of the matrix H is equal to the rank of the matrix X:

$$\sum_i h_{ii} = m = p+1.$$

Speaking more generally, if h_{ii} is large and y_i is aberrant, then the fitted value $\hat{y}_i$ will be determined mainly by y_i; moreover, the fitted value $\hat{y}_i$ will be pulled toward y_i and the fitted model may be seriously biased (distorted). Hoaglin and Welsch (1978) suggest identifying a leverage value h_{ii} as large one if:

$$h_{ii} \geqslant 2m/n. \tag{7}$$

For our data the cut-off point $2m/n$ established by (7) is equal to $28/149 = 0.1879$.

3.2. *Externally Studentized residuals*

Let $\hat{e}_i = y_i - \hat{y}_i$ be the residual that results from the fit. The variance of this residual is $(1-h_{ii})\sigma^2$, hence $\hat{e}_i/[(1-h_{ii})^{1/2}\sigma]$ has mean value 0 and variance 1. Usually, σ^2 is estimated by s^2, the residual variance:

$$\hat{\sigma}^2 = s^2 = (n-m)^{-1} \sum_{i=1}^{n} \hat{e}_i^2. \tag{8}$$

Let $s(i)$ be the standard deviation s computed from the data after deleting the i-th data vector; likewise let $\hat{\boldsymbol{\beta}}(i)$ be the vector of regression coefficients estimated from such deleted data, and let $X(i)$ be the matrix after deleting the i-th data vector (i.e. the i-th row of X). Let $\mathbf{x}_i$ be the i-th row of the design matrix X. So we have

$$t_i(i) = \frac{y_i - \mathbf{x}_i\hat{\beta}(i)}{s(i)[1+\mathbf{x}_i(X(i)'X(i))^{-1}\mathbf{x}_i']^{1/2}} = \frac{\hat{e}_i}{s(i)[1-h_{ii}]^{1/2}}. \tag{9}$$

In the statistical literature $t(i)$ described by the formula (9) is called an externally Studentized residual (Cook and Weisberg, 1980, Gibbons *et al.* 1987), deleted Studentized residual (Hocking, 1983), cross-validatory or jackknife residual (Atkinson, 1981). In his recent book Atkinson (1987) proposes to call it simply deletion residual. The updating and deletion formulas that underlie the computing of $t(i)$ can be found, among others, in Belsley *et al.* (1980) and Atkinson (1987). Under assumption of normality $N(0, \sigma^2 I)$ of the vector **e** appearing in formula (2) the variables $t(i)$ have a Student's t distribution with $n-p-1$ degrees of freedom. Belsley *et al.* (1980) propose to pay a special attention to the cases, for which

$$|t(i)| \geqslant 2.0. \tag{10}$$

A convenient way of identifying aberrant values of Y is plotting the deletion residuals against the fitted values of y.

3.3. DFFITS, the change in fitted value when deleting single individuals (items)

DFFITS is a statistics introduced by Belsley *et al.* (1980). It is defined as follows:

$$DFFITS(i) = (\hat{y}_i - \hat{y}_i(i))/[s(i)h_{ii}^{1/2}]. \tag{11}$$

In the numerator we have the change in fitted values that occurs when the i-th data vector is deleted. This difference is scaled by an estimate of the standard deviation of the fitted value. The formula (11) is equivalent to the following one:

$$DFFITS(i) = [h_{ii}/(1-h_{ii})]^{1/2} t_i(i). \tag{12}$$

It follows from (12) that $DFFITS(i)$ combines informations from h_{ii} (Eq. (5)) and $t(i)$ (Eq. (9)); $DFFITS(i)$ is large if either h_i or t_i is large. Belsley *et al.* (1980) recommend flagging $DFFITS_i$ as large one if:

$$|DFFITS(i)| \geqslant 2(p/n)^{1/2}. \tag{13}$$

For our data set the cut-off point established by (13) is equal to 0.6131.

3.4. DFBETAS, the change in the regression coefficients after deleting single data vectors

A scaled measure of the change in $\hat{\beta}_j$, $j = 0, \ldots, m$ was proposed by Belsley *et al.* (1980). It is called $DFBETAS_{ij}$ and is defined as follows:

$$DFBETAS_{ij} = [\hat{\beta}_j - \hat{\beta}_j(i)]/[s^2(i)\, c_{jj}]^{1/2}, \tag{14}$$

where c_{jj} is the j-th diagonal element of the matrix $C = (X'X)^{-1}$.

Several criteria have been proposed for stating whether a $DFBETAS$ is so large that it indicates an "influential" data vector. One criterion is the cut-off value equal to $2/\sqrt{n}$ (Belsley *et al.* 1980). The measure od $DFBETAS_{ij}$ expresses a scaled change of the j-th coefficient $\hat{\beta}_j$ evaluated from the whole data set after removing the i-th data vector from the data set.

In this paper we will employ another statistics defined as follows:

$$MDB(i) = \sum_{j=0}^{p} \left|\frac{\hat{\beta}_j - \hat{\beta}_j(i)}{\hat{\beta}_j}\right| 100. \qquad (15)$$

4. Analysis of the data

4.1. The values of the employed regression diagnostics

We consider the regression equations allowing to predict Fs or Fh taken in turn as the predicted variable Y. We carried out the calculations separately in the data sets E, C and A described in Chapter 1. In Tables 1 and 2 we show the values of the above introduced regression diagnostics (i.e. h_{ii}, $t(i)$, $DFFITS(i)$ and $MDB(i)$) evaluated for the Fs and Fh regressions. We show only the values for the selected data vectors exhibiting some features of unstability. The values of h_{ii}, $t(i)$ or $DFFITS(i)$ surpassing the appropriate cut-off points are marked by "+". Moreover, in these tables also the observed values of the predicted variables y_i are given together with the values of $\hat{y}_i(i)$. Thirteen data vectors (items) analysed in more detail by Bartkowiak and Jakimiec (1989) are marked by stars.

One can see from Tables 1 and 2 that in the data set E there are fifteen data vectors for which the values of the statistics h_{ii} surpass the cut-off point 0.1879. In particular, the highest value 0.97 occurs for the erroneous data vector no. 30 (discussed in more detail in the paper of Bartkowiak and Jakimiec, 1989). After correcting this data vector the new value h is equal to 0.06. Also the values of $DFFITS(i)$ and $MDB(i)$ are extremally high for this erroneous data vector, and after correction of this item the values diminish significantly. In the corrected data set C there remain still twelve items with high values of h_{ii}, i.e. there are still data vectors which could possibly influence considerably the regression. In most cases these data vectors were already revealed as outliers by means of other methods presented by Bartkowiak and Jakimiec (1988).

Comparing the values $t(i)$ calculated for the data sets E and C one can see that some of them are not diminished after removing the gross errors but even rised (e.g. for Fh the absolute value $|t(50)|$ rised from 1.66 to 2.14, and the absolute value $|t(75)|$ rised from 1.83 to 2.14). The existence of large values of h_{ii} (providing leverages for the fitted regressions) and simultaneously of relatively large values of $t(i)$ (showing large differences between the values y_i and $\hat{y}_i$ estimated independently, i.e. by the leaving-one-out method) allow us to suppose that the data vectors nos. 15 and 50 might be influential both for the regression of Fs and Fh. Generally, in the corrected data set C we find six

Table 1

Regression diagnostics evaluated in the data sets E, C, A for selected items. *Fs* considered as the predicted variable. * means that the item was discussed in the paper by Bartkowiak and Jakimiec (1988). + means that the diagnostic surpasses its cut-off value.

item		$\hat{y}_i$			h_{ii}			t(i)			DFFITS(i)			MDB(I)		
no.i	y_i	E	C	A	E	C	A	E	C	A	E	C	A	E	C	A
1*	1.97	2.04	1.98	2.08	0.27+	0.30+	0.35+	-0.20	-0.02	-0.32	-0.12	-0.01	-0.24	3.2	0.3	3.0
2	0.48	-0.10	-0.11	-0.10	0.04	0.05	0.05	1.36	1.40	1.42	0.29	0.31	0.32	3.5	3.5	3.7
8	0.99	0.64	0.73	0.73	0.20+	0.20+	0.20+	0.90	0.67	0.68	0.46	0.34	0.34	13.2	7.9	7.8
14	2.38	1.72	1.71	1.85	0.13	0.15	0.17	1.63	1.68	1.39	0.63+	0.70+	0.63+	19.7	15.5	11.9
15*	0.70	1.54	1.52	-	0.21+	0.25+	-	-2.19+	-2.22+	-	-1.14+	-1.28+	-	29.8	29.9	-
18	2.33	1.47	1.52	1.60	0.09	0.10	0.11	2.09+	1.99	1.84	0.67+	0.66+	0.66+	13.5	17.2	19.2
19	1.83	1.32	1.46	1.60	0.22+	0.31+	0.33+	1.32	1.01	0.67	0.70+	0.68+	0.47	20.8	18.2	10.1
22*	0.93	0.25	0.47	0.44	0.23+	0.07	0.07	1.77	1.09	1.20	0.96+	0.29	0.33	24.0	8.8	9.4
28*	1.26	1.71	1.46	1.58	0.12	0.26+	0.27+	-1.09	-0.54	-0.89	-0.40	-0.32	-0.54	14.5	6.6	9.8
30*	1.80■	1.68	0.93	0.94	0.97+	0.06	0.07	1.66	0.84	0.84	9.90+	0.22	0.22	68.8	6.4	6.9
35*	0.98	0.69	0.64	0.61	0.23+	0.07	0.08	0.74	0.82	0.92	0.40	0.23	0.27	19.0	8.0	9.8
50*	0.90	1.72	1.73	-	0.24+	0.27+	-	-2.17+	-2.28+	-	-1.21+	-1.39+	-	20.5	35.6	-
57*	2.06	1.91	2.15	2.28	0.30+	0.25+	0.27+	0.40	-0.24	-0.60	0.26	-0.14	-0.36	7.2	2.2	5.3
58*	1.95	2.04	2.16	2.29	0.25+	0.29+	0.31+	-0.24	-0.56	-0.97	-0.14	-0.36	-0.65+	3.2	7.9	13.3
59*	1.58	1.27	1.26	1.34	0.20+	0.40+	0.42+	0.78	0.94	0.73	0.39	0.77+	0.63+	13.4	15.7	12.0
61	1.26	0.48	0.43	0.45	0.08	0.08	0.08	1.86	2.01+	2.03+	0.55	0.60	0.61+	16.5	12.2	13.4
66	0.51	0.68	0.70	0.68	0.20+	0.20+	0.20+	-0.42	-0.48	-0.44	-0.21	-0.24	-0.22	6.6	6.4	6.5
68	-0.10	0.18	0.17	0.19	0.20+	0.20+	0.21+	-0.71	-0.70	-0.78	-0.36	-0.35	-0.41	11.4	10.9	14.6
75	1.66	1.02	0.95	1.03	0.14	0.14	0.15	1.57	1.78	1.63	0.63+	0.73+	0.68+	17.7	18.8	16.8
94	1.22	0.36	0.35	0.31	0.08	0.09	0.09	2.05+	2.13+	2.28+	0.59	0.65+	0.70+	15.1	7.8	17.6
101	1.04	-0.05	-0.04	-0.03	0.05	0.05	0.05	2.59+	2.62+	2.66+	0.58	0.59	0.62+	13.8	12.5	15.6
102	0.54	0.73	0.83	0.89	0.24+	0.26+	0.26+	-0.49	-0.77	-0.96	-0.27	-0.45	-0.57	7.5	11.3	12.4
104*	-0.10	-0.14	0.01	-0.02	0.29+	0.16	0.16	0.12	-0.27	-0.20	0.07	-0.11	-0.09	1.5	2.7	2.0
121*	0.00	0.61	0.27	0.24	0.07	0.08	0.08	-1.43	-0.64	-0.59	-0.38	-0.19	-0.17	10.5	4.0	3.7
133*	-0.10	0.48	0.49	0.50	0.07	0.07	0.07	-1.37	-1.41	-1.47	-0.36	-0.37	-0.39	11.2	7.7	8.9
140	1.23	0.30	0.29	0.26	0.06	0.06	0.06	2.20+	2.28+	2.41+	0.55	0.58	0.63+	16.6	15.3	16.7

■ y_{30} = 1.28 The values of y are the same in sets E, C, A

except for y_{30}, which equals in set E y = 1.80;

after correction it equals y = 1.28 both in set C and A.

Table 2

The description as in Table 1 but for the predicted variable *Fh*.

item		$\hat{y}_i$			h_{ii}			t(i)			DFFITS(i)			MDB(i)		
no.i	y_i	E	C	A	E	C	A	E	C	A	E	C	A	E	C	A
1*	2.89	2.77	3.10	3.28	0.27+	0.30+	0.35+	0.19	-0.36	-0.71	0.12	-0.24	-0.52	1.7	5.3	8.2
2	1.32	-0.13	-0.11	-0.09	0.04	0.05	0.05	2.04+	2.16+	2.17+	0.44	0.48	0.49	4.3	5.5	5.7
8	1.48	1.00	1.06	1.07	0.20+	0.20+	0.20+	0.72	0.68	0.68	0.37	0.34	0.34	7.8	9.7	10.9
14	3.62	2.58	2.59	2.76	0.13	0.15	0.17	1.51	1.63	1.41	0.58	0.67+	0.64+	9.6	15.8	14.6
15*	1.30	2.58	2.29	-	0.21+	0.25+	-	-1.98	-1.68	-	-1.03+	-0.97+	-	14.7	25.9	-
18	3.30	2.18	2.35	2.48	0.09	0.10	0.11	1.60	1.47	1.29	0.52	0.48	0.46	6.3	13.9	17.4
19	2.63	1.56	2.08	2.24	0.22+	0.31+	0.33+	1.66	0.95	0.70	0.88+	0.64+	0.49	11.4	22.1	13.4
22*	1.57	0.11	0.66	0.61	0.23+	0.07	0.07	2.29+	1.38	1.48	1.24+	0.37	0.40	17.2	10.6	13.8
28	1.79	3.28	2.26	2.41	0.12	0.26+	0.27+	-2.19+	-0.80	-1.07	-0.80+	-0.47	-0.65+	14.4	12.7	15.2
30*	15.00■	14.50	1.30	1.32	0.97+	0.06	0.07	4.39+	0.83	0.81	26.19+	0.21	0.21	57.3	7.5	9.8
35*	1.51	0.56	0.90	0.85	0.23+	0.07	0.08	1.47	0.92	1.03	0.80+	0.26	0.30	17.0	9.5	13.5
50*	1.48	2.54	2.72	-	0.24+	0.27+	-	-1.66	-2.14+	-	-0.93+	-1.30+	-	11.2	40.0	-
57*	3.36	3.04	3.43	3.62	0.30+	0.25+	0.27+	0.52	-0.12	-0.46	0.34	-0.07	-0.27	5.6	1.1	5.2
58*	3.37	3.56	3.40	3.61	0.25+	0.29+	0.31+	-0.29	-0.06	-0.42	-0.17	-0.04	-0.28	2.4	0.8	7.6
59*	2.42	2.86	2.05	2.13	0.20+	0.40+	0.42+	-0.66	0.70	0.56	-0.33	0.57	0.48	6.7	16.9	12.7
61	2.31	0.83	0.79	0.80	0.08	0.08	0.08	2.12+	2.34+	2.37+	0.63+	0.70+	0.71+	13.4	16.5	22.9
66	1.15	1.10	1.12	1.10	0.20+	0.20+	0.20+	0.07	0.05	0.09	0.04	0.02	0.04	0.8	0.7	1.6
68*	-0.10	0.41	0.35	0.66	0.20+	0.20+	0.20+	-0.77	-0.73	-0.76	-0.39	-0.37	-0.39	7.6	9.9	15.2
75	2.82	1.58	1.48	1.59	0.14	0.14	0.15	1.83	2.14+	2.00+	0.74+	0.88+	0.83+	13.1	20.0	21.6
94	1.71	0.79	0.67	0.62	0.08	0.09	0.09	1.30	1.59	1.70	0.37	0.49	0.53	6.7	14.6	15.9
101	2.00	-0.02	0.00	0.03	0.05	0.05	0.05	2.89+	3.07+	3.09+	0.64+	0.69+	0.72+	10.8	16.2	22.9
102	0.70	1.22	1.47	1.55	0.24+	0.26+	0.26+	-0.80	-1.31	-1.47	-0.45	-0.77+	-0.87+	7.8	16.9	17.7
104*	-0.10	-0.14	0.12	0.07	0.29+	0.16	0.16	0.06	-0.35	-0.27	0.04	-0.15	-0.12	0.6	4.0	3.4
121*	0.00	1.27	0.46	0.43	0.07	0.08	0.08	-1.79	-0.69	-0.66	-0.48	-0.20	-0.19	6.9	4.5	5.1
133*	-0.10	1.51	0.71	0.72	0.07	0.06	0.07	-2.30	-1.22	-1.25	-0.61+	-0.32	-0.34	13.9	8.3	10.8
140	1.88	0.47	0.43	0.40	0.06	0.06	0.06	2.00+	2.22+	2.29+	0.50	0.56	0.60	9.0	15.8	20.4

■ y_0= 1.85 The values of y are the same in sets E, C, A except for y_{30} ,which equals in set E y = 15.00; after correction it equals y = 1.85 both in set C and A.

values of $t(i)$ which exceed the cut-off point 2.0. The largest values are $t(101) = 2.62$ for Fs and $t(101) = 3.07$ for Fh. This does not seem to be especially contradictory to the Student's t distribution with 135 degrees of freedom for a sample size $n = 149$.

Looking at the values of the statistics $DFFITS(i)$ and $MDB(i)$ calculated for the data set C one can see that the largest values can be found for the data vectors nos. 15 and 50 for the regression predicting Fs and also Fh. The mean change of the regression coefficients exceeds 30% when removing the 50-th item and 20% when removing the 15-th item. In the data set E the highest and realy extreme values of MDB can be found for the erroneous data vector no. 30.

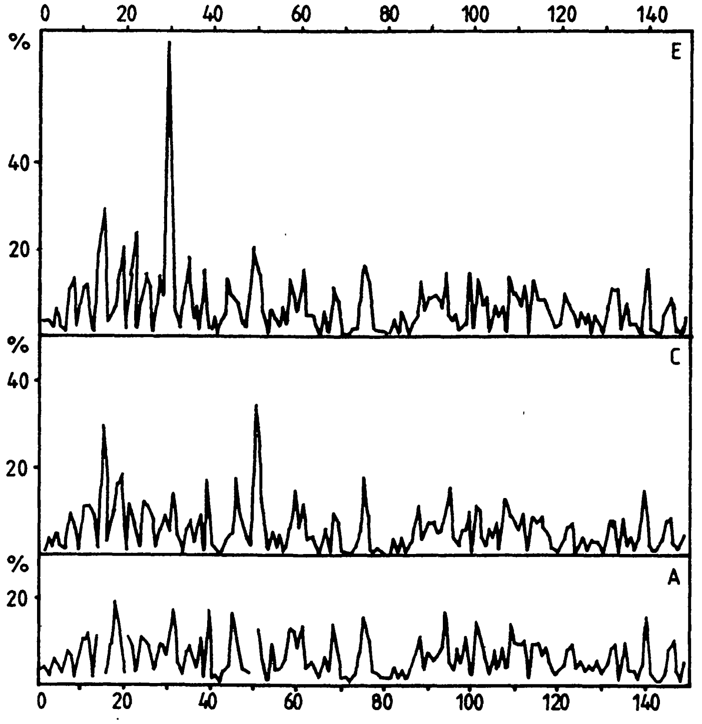

Fig. 1. Index plot of $MDB(i)$ put against the no. of the data vector, i. The values $MDB(i)$ are evaluated for the predicted variable Fs.

In Figures 1 and 2 the index plots of $MDB(i)$ put against i, the current number of the item, are shown. They were evaluated for the regression of Fs and Fh, respectively. The data vectors of big influence on the regression can be easily seen in these figures. First of all we see one such vector in the data set E (item no. 30), and two items in the set C (nos. 15 and 50). Excluding these two individuals from the set C we obtained a new data set A with $n = 147$ data vectors. Comparing the values calculated for the data sets C and A (shown in

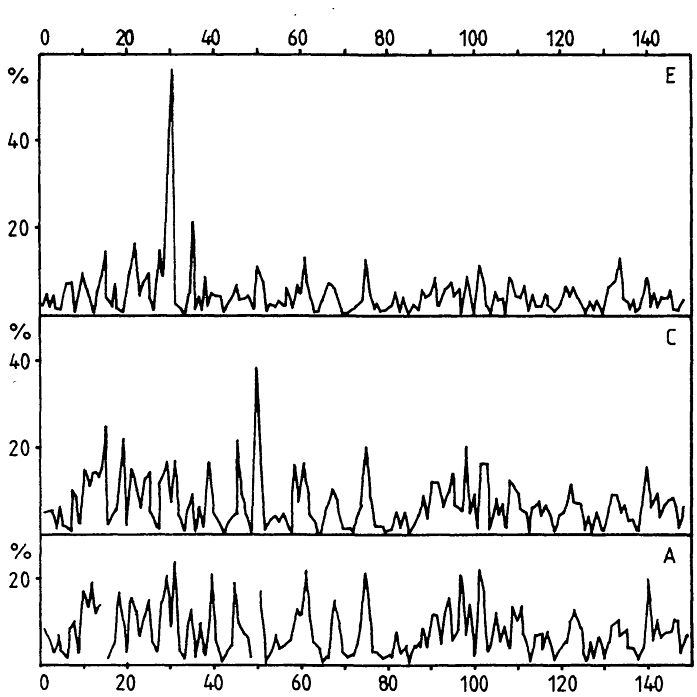

Fig. 2. The description as for Fig. 1 but for the predicted variable *Fh*.

Tables 1 and 2) one can see that the values of the analysed statistics (h_{ii}, $t(i)$, $DFFITS(i)$ and $MDB(i)$) did not change strongly after removing the two mentioned data vectors.

4.2. *Comparison of the regression coefficients evaluated from the data sets E, C and A*

In Tables 3 and 4 the estimated values of the regression coefficients b_j evaluated for the predicted variables *Fs* and *Fh*, are given, respectively. The

Table 3

Regression coefficients (and their standardized values) of the equation (1) for the predicted variable $y = Fs$. RR is the square of the multiple correlation coefficient.

j	data set E		data set C		data set A	
	b_j	$b_j/s(b_j)$	b_j	$b_j/s(b_j)$	b_j	$b_j/s(b_j)$
0	-1.253446		-1.213337		-1.143233	
1	0.161981	0.75	0.171860	0.81	0.171385	0.83
2	-0.106287	-0.67	-0.157187	-0.99	-0.158430	-1.02
3	0.336166	2.14	0.374059	1.95	0.373228	2.00
4	1.922023	2.30	1.853797	2.20	1.622157	1.97
5	0.017988	0.47	0.020081	0.54	0.035361	0.96
6	-0.029693	-0.49	-0.020583	-0.34	-0.018611	-0.32
7	0.181411	1.66	0.202519	1.87	0.208595	1.98
8	0.058965	0.26	0.163965	0.59	0.287151	1.04
9	-0.007320	-0.32	-0.010841	-0.47	-0.007475	-0.31
10	0.076948	3.15	0.071111	2.89	0.080814	2.83
11	0.476816	1.39	0.734600	1.84	0.770933	1.88
12	0.736378	1.97	-1.142173	-0.51	-1.298916	-0.57
13	-0.197967	-1.12	-0.351051	-1.47	-0.430363	-1.80
RR	0.55		0.56		0.59	

Table 4
The description as in Table 3 but for the predicted variable *Fh*.

j	data set E		data set C		data set A	
	b_j	$b_j/s(b_j)$	b_j	$b_j/s(b_j)$	b_j	$b_j/s(b_j)$
0	-1.740065		-1.558511		-1.450902	
1	0.220439	0.61	0.155334	0.46	0.152994	0.46
2	-0.237810	-0.89	-0.277597	-1.10	-0.281386	-1.14
3	0.676850	2.57	0.517804	1.71	0.512842	1.72
4	2.913094	2.08	2.954364	2.22	2.619828	1.99
5	0.059061	0.93	0.063059	1.07	0.082682	1.41
6	-0.095678	-0.94	-0.051559	-0.54	-0.044954	-0.48
7	0.246919	1.35	0.329253	1.92	0.334395	1.98
8	-0.398220	-1.05	0.179802	0.41	0.360375	0.82
9	-0.017767	-0.47	-0.014600	-0.40	-0.005906	-0.15
10	0.124202	3.04	0.120992	3.10	0.141495	3.11
11	1.021138	1.77	0.917640	1.45	0.906732	1.39
12	9.829225	15.74	-1.067274	-0.30	-1.637548	-0.45
13	-0.574908	-1.94	-0.425818	-1.13	-0.507999	-1.33
RR	0.78		0.53		0.55	

standardized values $b_j/s(b_j)$ shown also in these tables allow us to compare the values of these coefficients obtained for different sets of data. In lower part of the tables the appropriate values of the determination coefficient *RR* (an estimate of the square of the multiple correlation coefficient between the predicted variable $Y$and the set of predictors $X_1, X_2, \ldots, X_p$) are given. One can see the big differences between the values of regression coefficients obtained for the data sets E and C – especially for the coefficient b_{12}. These differences are due to the big error in the data vector no. 30 contained in the data set E. The differences in b_j between the data sets C and A are lower, but also noticeable.

4.3. *Using the values of the regression diagnostics for the differentiation of the data vectors*

Let us recall the meaning of the analysed statistics: Large value of a h_{ii} can (but not must) be connected with a big influence of the i-th data vector on the regression. A large value of $t(i)$ means a large deletion residual calculated as the difference between the observed value y_i and the expected value $\hat{y}_i(i)$ evaluated from a regression function with coefficients estimated from the data set after excluding the i-th data vector; In other words, $t(i)$ is large when the observed value y_i is far from the regression line. A large value of $DFFITS(i)$ means a big change in fitted values that results from the deletion of the i-th data vector. A large value of $MDB(i)$ means a big influence of the i-th data vector on the estimated values of the regression coefficients.

In Figures 3 and 4 we show the scatter diagrams of the values $\hat{y}$ put against the values y of the predicted variables Fs and Fh, respectively, evaluated for the data set C. We use the regression diagnostics for the differentiation of the data vectors. The vectors for which the value of $t(i) < 1.0$ and the value of $h_{ii} < 0.19$

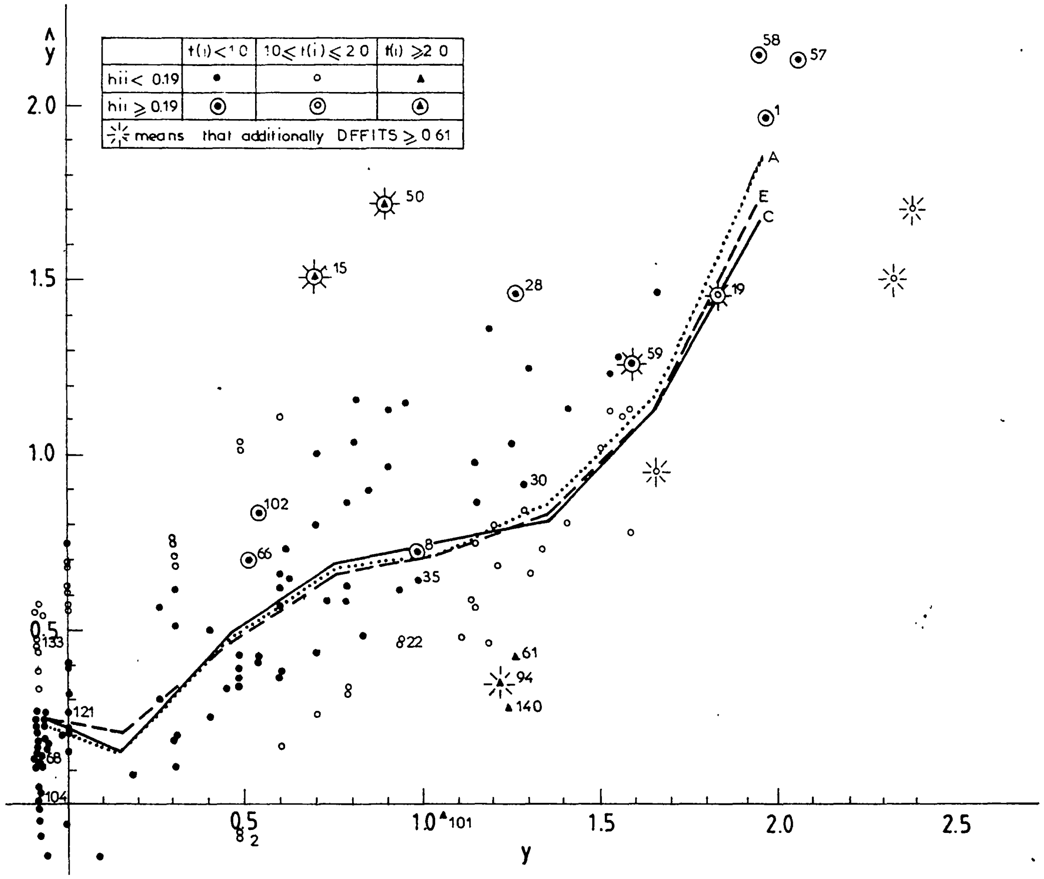

Fig. 3. Scatterdiagram of the pairs ($\hat{y}$, y), with *Fs* taken as the predicted variable, evaluated in the data set C. Curves denote the conditional regressions.

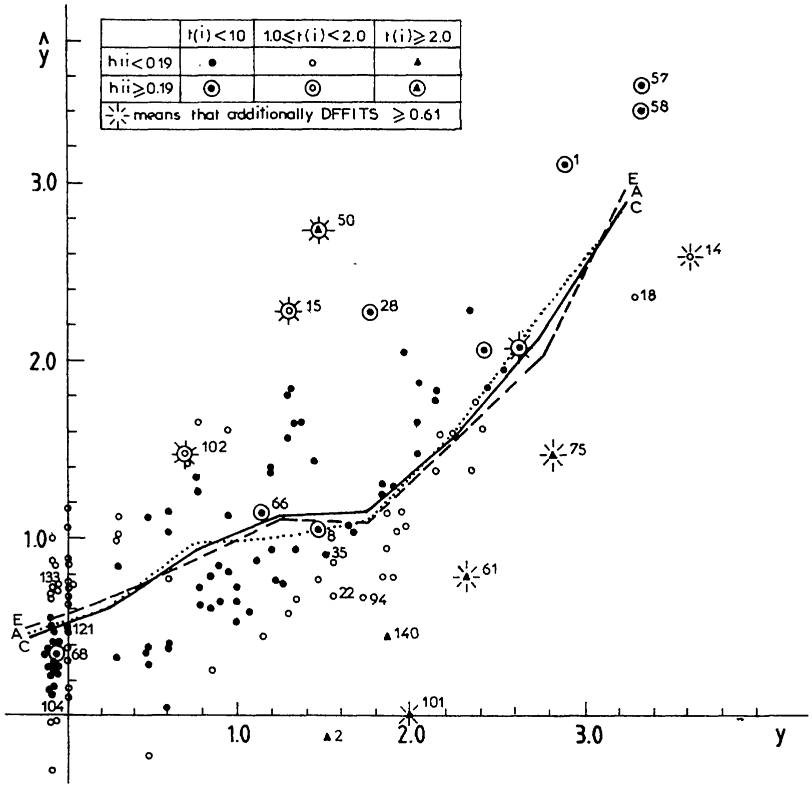

Fig. 4. The description as for Fig. 3 but for the predicted variable *Fh*.

are marked by black points. The vectors for which $1.0 \leqslant t(i) < 2.0$ and $h_{ii} < 0.19$ are marked by circles. The vectors for which $t(i) \geqslant 2.0$ and $h_{ii} < 0.19$ are marked by dark triangles. Each circled symbol (black point, circle or triangle) corresponds to the data vector for which $h_{ii} \geqslant 0.19$. Moreover, the data vectors for which $DFFITS(i) \geqslant 0.61$ are marked in Figures 3 and 4 by radial beams. The twenty four data wectors, for which the values of the regression diagnostics are collected in Tables 1 and 2, are denoted by numbers. The curves denote regression lines of the first kind and will be discussed in Chapter 5. One can see from Figures 3 and 4 that:

1. The data vectors revealing small values of both statistics h_{ii} and $t(i)$ are marked by black points. They make the essential part of the data vectors (56% for Fs and 60% for Fh). All they are located near the line $\hat{y} = y$ (e.g. items nos. 30, 35, 104 and 121). Rejection of one of such data vectors has virtually no effect on the estimated values of the regression coefficients. The mean value od $DFFITS$ evaluated for these data vectors (marked by black points) is about 0.12 only, so we might suppose that their influence on the regression is practically none.

2. The data vectors with small values of $t(i)$ and with the values of h_{ii} greater than the cut-off point are marked by circled points. They indicate possibly influential data vectors which could reveal either atypical interrelations between the variables X (e.g. items nos. 8, 28, 66, 68) or extremally large values of the variables X (e.g. items nos. 1, 57, 58). In our figures displaying the data set C, there are only a few of such data vectors. These points were already discovered and discussed in the former paper (Bartkowiak and Jakimiec, 1989). The mean value, about 0.28, of $DFFITS$ for these points is very moderate when compared to the cut-off value equal to 0.61.

3. Vectors revealing a small value of h_{ii} but a considerably larger value of the residuals ($1.0 \leqslant t(i) < 2.0$) are marked by small circles. They constitute a large part of the data (29% for Fs and 34% for Fh). The circles corresponding to these data vectors do not lie in Figures 3 and 4 on the periphery of the cluster of points-individuals but exhibit a specific pattern: for small values of the predicted variable (low X-ray flare activity) the values of $\hat{y}$ are greater than y (e.g. for item no. 133), whereas for high values of the predicted variable (high X-ray flare activity) the values of $\hat{y}$ are smaller than y (e.g. for items nos. 14, 18, 22). Their influence on the regression is stronger, the mean value of $DFFITS$ is for them about 0.39.

4. In our data set there are also a few data vectors (marked by radial beams) for which $DFFITS > 0.61$, what means a big impact on the fit of the regression model. These are data vectors for which either the value of h_{ii} is high ($h_{ii} > 0.19$) and the value of $t(i)$ is greater than 1.0 (these data vectors are marked by circled circles in Figures 3 and 4, e.g. item no. 19 for Fs and nos. 15, 102 for Fh) or $t(i)$ is greater than 2.0 (data vectors marked by triangles e.g. items nos. 15, 50, 61, 94, 101 and 140 for Fs and nos. 2, 50, 61, 75, 101, 140 for

Fh). The data vectors nos. 15 and 50 correspond to sunspot groups with a rapid decay of the X-ray flare activity, whereas the other vectors correspond to sunspot groups with a rapid increase of the flare activity. The impact of these data vectors on the regression seems to be very great. Especially the vectors nos. 15 and 50, for which also the values h_{ii} are high, seem to be realy influential (they are marked in Figures 3 and 4 by circled dark triangles).

4.4. Analysis of the dependence between $\hat{y}$ and y considered in the data sets E, C and A

Next we analyse the expected values $\hat{y}$ evaluated from the appropriate regression equations and compare them with the observed values y. We subdivided the range of the observed values of y into eight intervals and assumed the same intervals for $\hat{y}$. The results of the comparison between $\hat{y}$ and y were put together in the form of contingency tables. Two of them (for $Y = Fs$ and $Y = Fh$, evaluated in the data set C) are shown in Table 5. In these tables

Table 5

Contingency tables for ($\hat{y}$, y) showing the accordance of the observed and expected values of y evaluated from the regression equation estimated from the data. Presented results are obtained from the data set C.

interval boundaries	observed values y=Fs	calculated values, $\hat{y}$								Total
		1	2	3	4	5	6	7	8	
	1	5	26	13	5	0	0	0	0	49
0.0										
	2	1	2	1	0	0	0	0	0	4
0.3										
	3	2	4	9	7	2	0	0	0	24
0.6										
	4	0	2	9	7	5	0	1	0	24
0.9										
	5	1	0	6	5	4	1	1	0	18
1.2										
	6	0	1	2	8	3	2	0	0	14
1.5										
	7	0	0	0	1	5	4	0	0	10
1.8										
	8	0	0	0	0	0	1	2	3	6
Total		9	35	40	31	19	8	4	3	149

interval boundaries	observed values y=Fh	calculated values, $\hat{y}$								Total
		1	2	3	4	5	6	7	8	
	1	4	26	17	3	0	0	0	0	50
0.0										
	2	1	4	2	3	0	0	0	0	10
0.5										
	3	0	4	8	7	2	0	0	0	21
1.0										
	4	1	1	12	5	5	1	1	0	26
1.5										
	5	0	1	7	9	0	2	0	0	19
2.0										
	6	0	1	1	3	8	2	0	0	15
2.5										
	7	0	0	0	1	1	1	0	1	4
3.0										
	8	0	0	0	0	0	1	1	2	4
Total		6	37	47	31	16	7	2	3	149

also the assumed interval boundaries are given. The diagonal elements are underlined. They correspond to the cases, for which the observed and expected values are very close to each other. They constitute 21% and 17% of total cases for *Fs* and *Fh*, respectively. Let us emphasize that these contingency tables were obtained from the same data sets for which the regression coefficients have been estimated. In the control tables one can also see some non-empty counts far from the diagonal – they comprise the atypical data vectors marked by dark triangles in Figures 3 and 4.

Table 6

Conditional means $\hat{y}_{cond}$ evaluated in 8 intervals of the range of y observed.

predicted variable	$y_{observed}$ – mean of interval	conditional mean $\hat{y}_{cond}$		
		data set E	data set C	data set A
Fs	-0.15	0.27	0.26	0.25
	0.15	0.21	0.15	0.15
	0.45	0.47	0.49	0.47
	0.75	0.66	0.69	0.63
	1.05	0.72	0.75	0.71
	1.35	0.84	0.81	0.86
	1.65	1.14	1.14	1.17
	1.95	1.69	1.75	1.85
Fh	-0.25	0.50	0.44	0.45
	0.25	0.65	0.60	0.60
	0.75	0.87	0.92	0.96
	1.25	1.10	1.12	1.00
	1.75	1.09	1.12	1.09
	2.25	1.52	1.55	1.58
	2.75	2.00	2.12	2.25
	3.25	2.95	2.88	2.88

In order to compare the interdependence between $\hat{y}$ and y in the three data sets E, C and A we evaluated the conditional regression of $\hat{y}$ in the given intervals of y (the regression of the first kind). For each interval of y we evaluated as the conditional mean $\hat{y}_{cond}$ the average of all $\hat{y}$ values falling into this interval. These conditional means are given in Table 6. From these values the conditional regression (regression of the first kind) were constructed. They are shown in Figures 3 and 4. One can see from the Table 6 and from Figures 3 and 4 that in spite of the differences in the values of the regression coefficients the constructed conditional regressions differ in principle neither for E and C nor for C and A data sets.

5. Discussion

In the short-term predictions of solar flare activity the building of an appropriate model is based on the assumption that the sunspot group does not change strongly from day to day, in other words, that the situation is rather stable. In fact, while many sunspot features do not change strongly from day to day (e.g. the sunspot group area or the magnetic strength), the local magnetic field configuration can change very quickly, even during several hours. Therefore the physical conditions which are favourable for flare activity can appear in a time shorter than 24 hours and also they can disappear quickly. It is somehow amazing that despite of this changing situation the variables characterizing flare activity of the sunspot group are the best predictors as it is emphasized e.g. by Sawyer *et al.* (1986). The estimated regression function should reflect true physical relations between the sunspot group characteristics on the given day and the flare activity on the next day. Such true relation can be established from the mean part of the data sets, after removing off the singular, atypical data vectors which perturb the regressions. In this paper we have described and applied some statistics (regression diagnostics) allowing to identify such atypical data vectors. Some of these vectors, for our data, characterize sunspot groups with unstable situation (rapidly changing flare activity).

One can see from the analysis that the differences between the predictions based on the data sets E, C and A are practicaly none. The differences in the regression coefficients are not very high. It is very encouraging that the predicting functions despite of some big errors or atypical data vectors occurring in the data are sufficiently stable, i.e. the predicting function is not disturbed strongly by singular data vectors even if the size of the training data set with $p = 13$ variables is only about 150. The problem of an extrapolation of the predicting function needs a further exploration, especially when in the new data set some atypical data vectors are encountered.

Finishing this discussion we would like to emphasize the fact that the contingency tables reveal doubtless a pattern of asymmetry. This fact indicates that the assumed regression model does not fit ideally into our data. Maybe, another mathematical model would give a better fit or, maybe, we should change fundamentally the model. Some suggestions for the construction of more sophisticated models can be found in the book by C. Sawyer *et al.* (1986). This fact needs further exploration.

REFERENCES

Atkinson, A. C., 1981, *Biometrika*, **68**, No. 1, 13.

— 1987, *Plots, Transformations and Regression. An Introduction to Graphical Methods of Diagnostic Regression Analysis*, IInd ed. (Oxford, University Press).

Bartkowiak, A. and Jakimiec, M., 1989, *Acta Astr.*, **39**, 85.

Belsley, D. A., Kuh, E., and Welsch, R. E., 1980, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (New York, Wiley).

Chatterjee, S. and Hadi, A. S., 1987, *Statistical Science*, **1**, 379.

Cook, R. D. and Weisberg, S., 1980, *Technometrics*, **22**, 495.

Gibbons, D. F., McDonald, G. C., and Gunst, R. F., 1987, *J. of Naval Research*, **34**, 109.

Hirman, J. W., Neidig, D. F., Seagraves, P. H., Flowers, W. E., and Wiborg, P. H., 1980, in *Sol.-Terres. Pred. Proc.*, vol. **3**, ed. R. F. Donnelly, C-64.

Hoaglin, D. C. and Welsch, R. E., 1978, *The American Statistician*, **32**, 17.

Hocking, P. R., 1983, *Technometrics*, **25**, 219.

Jakimiec, M. and Wasiucionek, J., 1980, in *Sol.-Terres. Pred. Proc.*, vol. **3**, ed. R. F. Donnelly, C-54.

Neidig, D. F., Wiborg, P. H., Seagraves, P. H., Hirman, J. W., and Flowers, W. E., 1986, in *Sol.-Terres. Pred. Proc.*, eds. P. A. Simon, G. Heckman and M. A. Shea, p. 300.

Sawyer. C., Warwick, J. W., and Dennett, J. T., 1986, *Solar Flare Prediction* (Boulder: Colorado Associated University Press).