ACTA ASTRONOMICA
Vol. 40 (1990) pp. 169–181

Short-term Predictions of Flare Activity Using α-Trimmed Regression Method

by

A. Bartkowiak

Institute of Computer Science, University of Wrocław, Poland

M. Jakimiec

Astronomical Institute, University of Wrocław, Poland

Received September 16, 1989

ABSTRACT

We consider data characterizing D, E, F sunspot groups in the decay phase. We construct, using $p = 13$ variables, a linear regression function allowing to predict the total X-ray flare flux for the next day from the characteristics recorded the given day. A robust α-trimmed regression was applied. This procedure permitted us to subdivide our data set into two parts. The first part, a major one, describes sunspot groups with slowly changing flaring process, for which the predictions of flare activity for the next day can be performed rather satisfactorily. The second part of the data, a smaller one, describes sunspot groups with sudden unexpected and unpredictable changes of flare activity for the next day.

1. Introduction

Short term predictions are often based on a regression equation of the type:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p + e, \tag{1}$$

where the predicted variable y is expressed as a linear function of p explanatory (predicting) variables $x_1, x_2, \ldots, x_p$ and an error term e. Such a regression model was, for example, considered by Bartkowiak and Jakimiec (1986). Their aim was to obtain a regression function allowing to predict flare activity (y) for the next day using some characteristics $x_1, x_2, \ldots, x_p$ describing the active region the given day. They achieved the aim employing a linear regression function (1) in which the parameters $b_0, b_1, \ldots, b_p$ were estimated by the least squares of errors method (LSE). The predictions obtained in this way were quite reasonable, what means that for the bulk of the data the discrepancy between the predicted and actually observed flare activity was not very high. However, the differences between the predicted and actually observed values displayed a decidedly non-symmetric pattern, and moreover, for some particular data vectors the authors obtained a very large discrepancy. We are confronted with the following problems:

(i) To find out whether the regression equation (1) constructed on the base of some empirical data is stable, *i.e.* whether it may remain almost the same after removing some data vectors from these data.

(ii) To investigate whether some other, more modern as the LSE methods, called robust methods, might yield a "better" predicting algorithm. By a "better" algorithm we mean a more appropriate and more suitable to the main part of the considered data points.

The problems described in (i) and (ii) are dealing with Eq. (1) and with estimation of its parameters. We might put another question:

(iii) Perhaps, other regression model than that given by Eq. (1) would be more suitable. There was some evidence that the assumed linear regression function (1), although very simple and easy to calculate, is somehow inadequate to our data and we can consider it to be only an approximation of the true relationship between the predicted and predicting variables.

In the present paper we take under consideration mainly the points (i) and (ii). We will show also that the assumed model (1) is not fully adequate for our data.

2. Investigation of the stability of the regression equation

We are likely to say that the regression equation is stable, if the regression will remain to be almost the same after removing from the training data set an arbitrary small subset of data vectors. We can consider the problem in a more precise way employing some special statistics, called regression diagnostics (see *e.g.* Atkinson,1987).

Investigating the problem of short-term flare activity prediction Jakimiec and Bartkowiak (1989) used four such regression diagnostics:

(a) diagonal elements of the hat matrix - pointing to leverage points in the regression,

(b) studentized residuals - pointing to data vectors which give relatively large residuals,

(c) DFFITS - a statistics, pointing to single data points, which, when removed from the process of estimation, give a considerable change in the fit of the regression, and

(d) DFBETAS - a statistics, pointing to single data points which, when removed from the process of estimation, give a considerable change in the coefficients of the regression equation.

Definitions and exact formulae for these statistics can be found *e.g.* in the mentioned paper by Jakimiec and Bartkowiak (1989). They considered $p = 13$ characteristics describing sunspot groups in the decay phase of their evolution (data for 1979 were gathered from Solar Geophysical Data). The predicted variables y were Fs and Fh, *i.e.* the daily sums of the X-ray flare fluxes in the wavelength intervals 1–8 Å (Fs) and 0.5–4 Å (Fh), respectively. Applying the statistics (a) – (d) described above they found several really influential data vectors which had great impact on the fit or on the estimated coefficients of the regression equation. These influential data vectors are related to the sunspot groups with either a very strong flare activity or a great, sudden change of activity from day to day. However, considering conditional regression of $\hat{y}$ in the given intervals of y (*i.e.* the regression of the first kind of the predicted values $\hat{y}$ on the observed values y) they found that, in spite of the differences in the values of the regression coefficients, the constructed conditional regressions are not different in principle, when obtained for all data vectors, and also for a reduced data set (after removing some influential data vectors). It means that the predicting algorithm is sufficiently stable in spite of atypical vectors contained in the training data set.

3. Robust regression

It can happen that in the training data set used in the estimation process some atypical, wild data vectors strikingly different than the others occur. Such wild data vectors can be damaging (very destructive) for the estimated values of the parameters. To find out which data vectors are influential for the estimated regression, one can compute regression diagnostics, *e.g.* those described in Section 2 of the paper. However, the computation process is somehow troublesome, because one investigates usually at one step the impact or importance of only one data vector. So, one must perform the computation in $n + 1$ steps (where n is the number of data vectors), n times removing only one data vector from the set possibly containing other influential data vectors.

To avoid this handicap, other methods of estimation were developed with the aim to make the estimated values to be resistant against atypical data vectors occurring in the data. We imagine that the whole considered data set is a mixture of some basic, essential data, for which the estimation process is carried out, and of a small contamination connected with atypical data points or simply errors. This contamination should not influence the estimates of the parameters expected to be proper for the main part of the data.

It has long be known that the LSE estimates are very sensitive to contaminations of the basic data. Therefore, to get protection against such contaminations, other estimation methods were developed. They are called robust methods. For a review of these methods see *e.g.* the monographs by Huber (1981) or Hampel *et al.* (1986). A general idea of these methods is to evaluate a kind of weighted estimates giving less weight to the data vectors which, for some reasons, are suspected to be "wild", not belonging to the main bulk of the data set.

We have chosen for our investigations the so called α-trimmed LSE regression. It seems to have good properties, what was stated in some simulation experiments (Antoch *et al.* 1984). Now we present in short the main features of the α-trimmed LSE regression:
Let r_i be the difference between the recorded (y_i) and predicted $(y_i^{(\alpha)} = b_0^{(\alpha)} + b_1^{(\alpha)}x_{i1} + \ldots + b_p^{(\alpha)}x_{ip})$ value of the considered variable y, for $i = 1, 2, \ldots, n$ (where n is the number of considered data vectors). Let $0 < \alpha < 1$ be a fixed real value. Find $b_0^{(\alpha)}, b_1^{(\alpha)}, \ldots, b_p^{(\alpha)}$ such, that the sum of very simply weighted residuals r_i is minimized:

$$\left\{b_0^{(\alpha)}, b_1^{(\alpha)}, \ldots, b_p^{(\alpha)}\right\} = \arg\left\{\min_{b_0, b_1, \ldots, b_p} \sum_{i=1}^{n} w_i r_i\right\}, \tag{2}$$

where

$$\begin{aligned} r_i &= y_i - b_0^{(\alpha)} - b_1^{(\alpha)}x_{i1} - \ldots - b_p^{(\alpha)}x_{ip}, \\ w_i &= \begin{cases} \alpha & \text{if } r_i > 0, \\ \alpha - 1 & \text{if } r_i < 0. \end{cases} \end{aligned} \tag{3}$$

It is advised for the practice to take for α such values as 0.05, 0.10, 0.15. Computing the coefficients $b_0^{(\alpha_1)}, b_1^{(\alpha_1)}, \ldots, b_p^{(\alpha_1)}$ for a given $\alpha = \alpha_1 < 0.5$ we obtain such regression equation which trimmes off $[\alpha_1 n]$ data vectors with big negative values of the residuals. Repeating this procedure with $\alpha = \alpha_2 = 1 - \alpha_1$, we get a regression with coefficients $b_0^{(\alpha_2)}, b_1^{(\alpha_2)}, \ldots, b_p^{(\alpha_2)}$ which trimmes off $[\alpha_2 n]$ data vectors with big positive values of the residuals (Fig. 1 illustrates this process). It is not always possible to obtain in the trimming process exactly an amount $[\alpha_1 n]$ or $[\alpha_2 n]$ data vectors satysfying the Eq. (3). The regression obtained from Eq. (2) should pass exactly through $p + 1$ data points. These points are not trimmed off from the data. In both cases, the method trimmes off points yielding large discrepancies between the observed values (y_i), and those calculated from the estimated regression function $(y_i^{(\alpha)})$. We reckon the remaining data set for homogeneous one. This reduced data set is used for computation of an ordinary LSE regression,

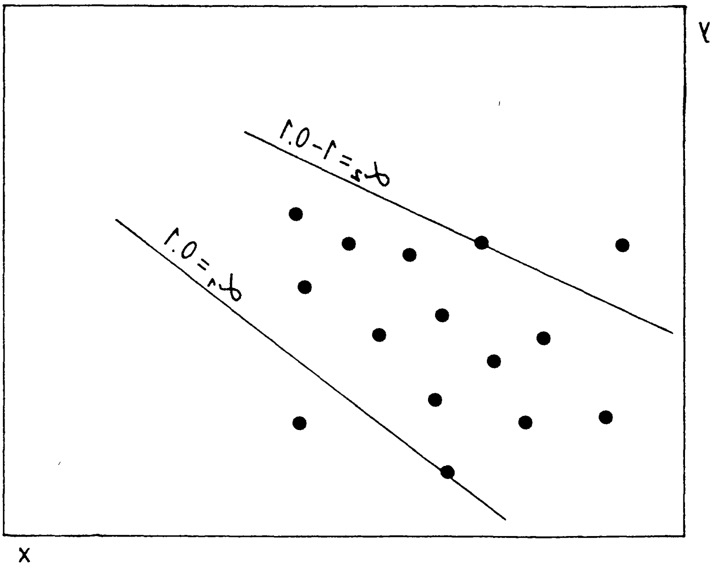

Fig. 1. Trimming off an $\alpha = 0.1$ part of points from the "bottom" and from the "top" of the (x,y) data cloud.

which is called now "α-trimmed LSE regression". The coefficients of this regression will be denoted as $b_0^{(\alpha)}, b_1^{(\alpha)}, \ldots, b_p^{(\alpha)}$

We used in our computations an algorithm developed by Antoch and described by Antoch *et al.* (1984). This algorithm with $\alpha = 0.10$ was applied to the same data which was previously considered by Jakimiec and Bartkowiak (1989). We remind that these data comprise $p = 13$ sunspot group characteristics employed as predicting variables, and two predicted variables ($y = Fs$ and $y = Fh$).

In Table 1 and Table 2 we show the estimates of the parameters $(b_0, b_1, \ldots, b_p)$ obtained by use of ordinary LSE method (denoted $\hat{b}_j$), and by use of the α-trimmed LSE method (denoted $b_j^{(\alpha)}$). Together with the appropriate values of b_j $(j = 0, 1, \ldots, 13)$ we show also the values t_j defined as follows:

$$t_j = \frac{b_j}{(Var(b_j))^{1/2}}. \tag{4}$$

In the formula above b_j stands for $\hat{b}_j$ or $b_j^{(\alpha)}$. From the values $\hat{b}_j$, $b_j^{(\alpha)}$ and t_j shown in Table 1 (prediction of Fs) and in Table 2 (prediction of Fh) we state that the difference between the estimates $\hat{b}_j$ (obtained by the LSE method) and $b_j^{(\alpha)}$ (obtained by the α-trimmed method) is not very large. From the t_j values, indicating the statistical significance of the j-th predicting variable in the regression equation, one can state, that the t-values obtained from the α-trimmed regression are for several predictors much more pronounced than those calculated from the LSE regression. Those of them, which are surpassing the value $t = 2.0$ are marked in Tables 1 and 2 by a "x" sign. Also R^2, the square of the multiple correlation coefficient, is

Table 1

Coefficients of the regression function obtained for the predicted variable Fs by use of the LSE and α-trimmed methods, together with the corresponding t values. R^2 is the square of the multiple correlation coefficient.

j	LSE method		α-trimmed method	
	$\hat{b}_j$	t_j	$\hat{b}_j^{(\alpha)}$	t_j
0	-1.2304	--	-1.3595	--
1	0.1442	0.67	0.1769	0.92
2	-0.0923	-0.58	-0.1497	-1.07
3	0.2472	1.50	0.2882	1.56
4	1.9950	2.41	2.1601	2.95x
5	0.0157	0.42	0.0317	0.94
6	-0.0171	-0.28	-0.0015	-0.03
7	0.2668	1.89	0.2432	2.52x
8	0.2233	0.92	0.3047	1.49
9	-0.0022	-0.10	0.0027	0.14
10	0.0783	3.23	0.0861	3.67x
11	0.2732	0.76	0.1230	0.39
12	-2.4840	-1.31	-3.5364	-2.12x
13	-0.0430	-0.22	0.0214	0.12
	R^2 = 0.5510	n = 149	R^2 = 0.6459	n = 133

higher for the α-trimmed method, both for Fs and for Fh. From the t_j values, given in Tables 1 and 2, we can see that several predicting variables (x_4 – the calcium plage intensity, x_7 – the magnetic field index, x_{10} – the number of stronger flares, and x_{12} – the hardness index, Fh/Fs) are much more pronounced than the others, *i.e.* they are indicated to be important predictors in the regression equation. The chosen subset of the predictors is the same for the both predicted variables, *i.e.* Fs and Fh.

To visualize, and to examine in more details the quality of the prediction we constructed scatterdiagrams of the values $(y_i, y_i^{(\alpha)})$, for $i = 1, 2, \ldots, n$, with $y_i^{(\alpha)}$ calculated from the α-trimmed LSE regression. These scatterdiagrams for Fs and Fh are shown in Figs. 2 and 3, respectively. The circled points correspond to the data vectors, which were trimmed off in the estimation process. The line corresponds to the conditional regression

Table 2

Coefficients of the regression function obtained for the predicted variable Fh by use of the LSE and α-trimmed methods, together with the corresponding t values. R^2 is the square of the multiple correlation coefficient.

	LSE method		α-trimmed method	
j	$\hat{b}_j$	t_j	$\hat{b}_j^{(\alpha)}$	t_j
0	-1.5722	--	-2.0423	--
1	0.1219	0.36	0.2064	0.72
2	-0.1795	-0.72	-0.1984	-0.92
3	0.3098	1.19	0.2640	1.18
4	3.1427	2.41	3.8093	3.40x
5	0.0567	0.96	0.0513	1.00
6	-0.0467	-0.49	-0.0564	-0.69
7	0.3404	1.97	0.4554	2.91x
8	0.2667	0.69	0.3245	1.01
9	-0.0012	-0.03	0.0039	0.13
10	0.1312	3.42	0.1374	3.92x
11	0.1994	0.35	-0.0220	-0.05
12	-3.3251	-1.11	-4.8539	-1.89
13	0.0629	0.20	0.2319	0.84
	R^2 = 0.5231	n = 149	R^2 = 0.6871	n = 129

line of $y^{(\alpha)}$ on y. Analogous scatterdiagrams were constructed by Jakimiec and Bartkowiak (1989) for the values ($\hat{y}_i$) calculated from the ordinary LSE regression. Obtained by them conditional regression lines of $\hat{y}$ on y are marked in Figs. 2 and 3 by the dashed lines. We do not see a big difference between the scatterdiagrams for the values calculated from the α-trimmed and from the LSE regression. There are some, nor very large, differences between the corresponding conditional regressions.

Dividing the values of y and $y^{(\alpha)}$ into 8 classes with class boundaries: (0.0, 0.3, 0.6, 0.9, 1.2, 1.5 and 1.8) for Fs, and (0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0) for Fh, we constructed contingency tables (Table 3) of counts n_{kl}, where k and l are the indices of the classes, into which the observed (y) and predicted ($y^{(\alpha)}$) values were accounted, respectively. As previously, the values of $y^{(\alpha)}$ were calculated from the α-trimmed regression. The trimmed off data vectors are marked by dots. In Table 3 one can see that the pattern

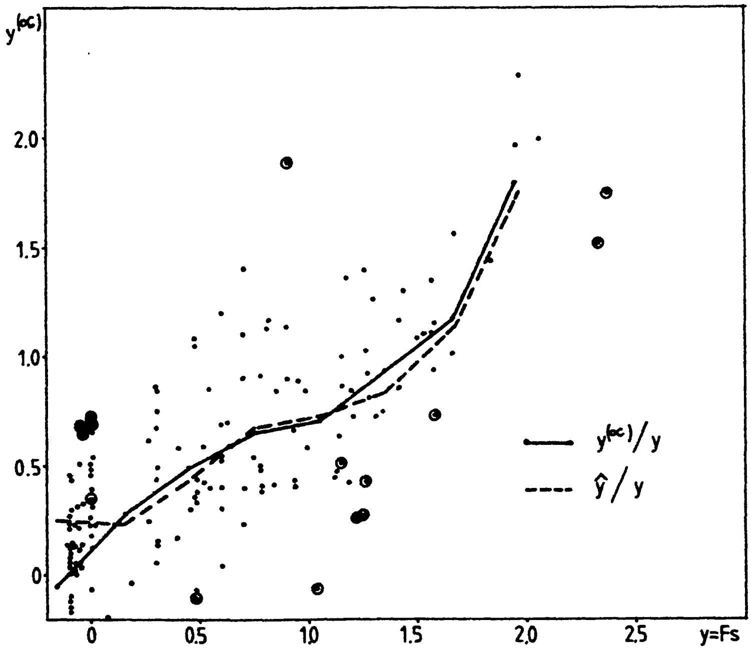

Fig. 2. Scatterdiagram of observed and predicted from the α-trimmed regression function values $(y, y^{(\alpha)})$ of the predicted variable $y = Fs$. The line denotes conditional regression line of $y^{(\alpha)}$ upon y. Circled points denote the data vectors trimmed off by the robust method.

of the counts is asymmetric with respect to the diagonal. The asymmetry effect, (discussed *e.g.* by Jakimiec and Wanke-Jakubowska, 1988), consist in overestimation of low flare activity and in underestimation of strong flare activity.

This effect can be seen more clearly in diagrams, exhibiting the difference $d = y^{(\alpha)} - y$ versus the values y, shown in Figs. 4 and 5 for the predicted variable Fs and Fh, respectively. In these diagrams one can see, that for smaller values of y the difference d tends to be rather positive (overestimation), while for larger values of y the corresponding differences are rather negative (underestimation). From this fact we infer that the assumed model is not adequate for our data (we will return to this point in the end of our paper). Moreover, it is of interest to know which data vectors were trimmed off by the α-trimmed regression. For the predicted variable Fs the algorithm has trimmed off 16 data vectors, and 6 of them were previously discovered by Jakimiec and Bartkowiak (1989) as influential (atypical) data vectors. Similarly, for the predicted variable Fh the algorithm has trimmed off 20 data vectors, and 6 of them were previously discovered as influential points. Not all, revealed previously as influential data points when considering regression diagnostics, were now trimmed off by the robust regression. The data vectors trimmed off by the α-trimmed regression are marked in Figs. 2–5 by circles. Their position in the contingency tables given in Table 3, is

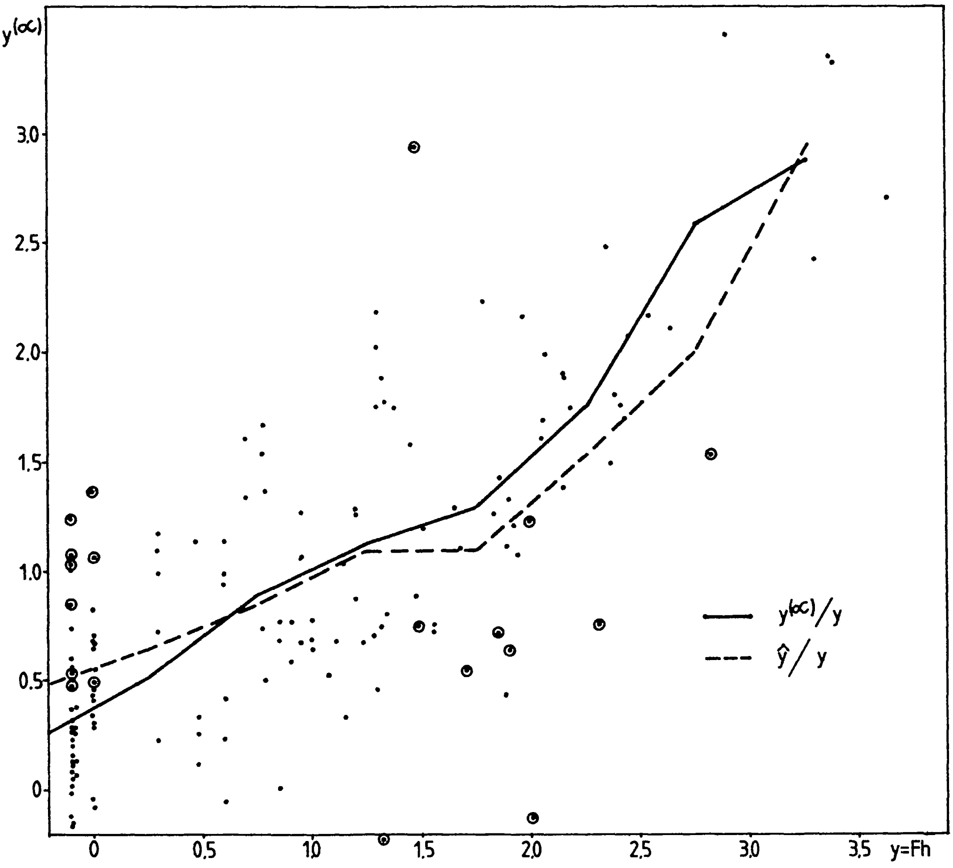

Fig. 3. Like Fig. 2, for the predicted variable $y = Fh$.

denoted by dots. One can see that a large part of the trimmed off points is located at rather extremly distant positions, far from the bulk of the data vectors.

4. Conclusions

The employed robust α-trimmed regression permitted to identify data points yielding large residuals. A large residual means that the flare activity appearing next day cannot be predicted satisfactorily from the characteristics of the sunspot group observed the given day. Most of the trimmed off data vectors are related to sunspot groups with sudden change of flare activity, *e.g.* related to a newly emerged magnetic flux. So, the bad prediction may be due to drastic, sudden changes of the interrelations of the active region characteristics. However, as it was found by Jakimiec and Bartkowiak (this issue), removing single, atypical data vectors from the data set, no significant change of the interrelation structure (described by common factors) of the characteristics is stated.

Jakimiec and Wanke-Jakubowska (1988) have removed from the training

Table 3

Contingency tables of n_{kl}, the counts of data vectors belonging to the k-th class of y (the observed values) and to the l-th interval of $y^{(\alpha)}$ (the predicted values). Dots indicate data vectors which were trimmed off by the robust regression algorithm.

observed	calculated, $y^{(\alpha)}$								
$y = Fs$	1	2	3	4	5	6	7	8	Total
1	4	21	6	2	0	0	0	0	33
2	3	6	7	4	0	0	0	0	20
3	2	4	11	5	2	0	0	0	24
4	0	2	12	3	6	1	0	0	24
5	1	0	7	5	3	1	0	1	18
6	0	2	1	6	3	2	0	0	14
7	0	0	0	1	6	2	1	0	10
8	0	0	0	0	0	1	2	3	6
Total	10	35	44	26	20	7	3	4	149

observed	calculated, $y^{(\alpha)}$								
y = Fh	1	2	3	4	5	6	7	8	Total
1	5	18	6	4	0	0	0	0	33
2	3	11	8	5	0	0	0	0	27
3	1	3	9	5	3	0	0	0	21
4	1	2	12	3	5	2	1	0	26
5	0	1	6	10	0	2	0	0	19
6	1	0	1	2	9	2	0	0	15
7	0	0	0	0	1	2	0	1	4
8	0	0	0	0	0	1	1	2	4
Total	11	35	42	29	18	9	2	3	149

data set these data vectors, which are related to the sunspot groups revealing sudden increase of flare activity. They found that both, the correlations between the characteristics and the predicting function are changed for this modified data set. In this paper, we applied the α-trimmed LSE method, which automatically removes from the data set the data vectors related to sudden rise or to sudden drop of flare activity. We found some, not very large, differences between the predicting functions obtained by use of the

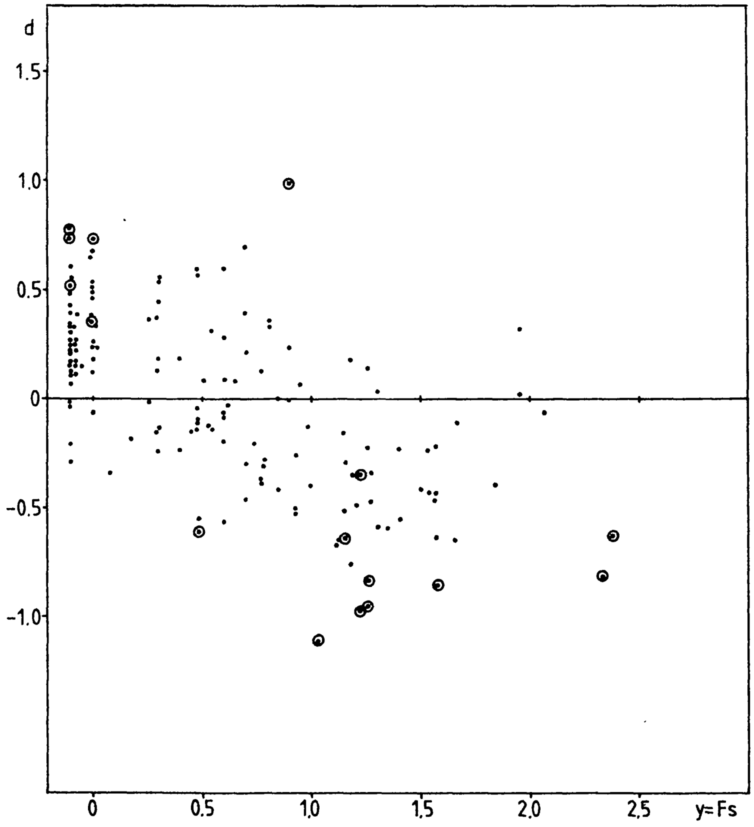

Fig. 4. The difference $d = y^{(\alpha)} - y$ between the predicted and observed values of the predicted variable $y = Fs$, versus y. The values of $y^{(\alpha)}$ are estimated from the α-trimmed regression with $\alpha = 0.1$.

standard LSE and the α-trimmed LSE methods. This fact seems to confirm the conclusion that the predicting algorithm is rather stable.

The α-trimmed method enables us to estimate the predicting function for the data describing a slowly changing process of flare activity. This process seems to be defined in principle by the evolution of active regions. If we would apply some other regression model, may be non linear, this process might be sufficiently good predictable. Instead, the process defined by the sudden changes of flare activity, superposed on the slowly changing flare activity process, is unpredictable as yet.

In a forthcoming work we want to consider the problem of constructing short-term predictions of flare activity using another form of robust regression, so called weighted regression with Huber weights. We will compare the performance of the α-trimmed LSE method and the weighted regression with Huber weights.

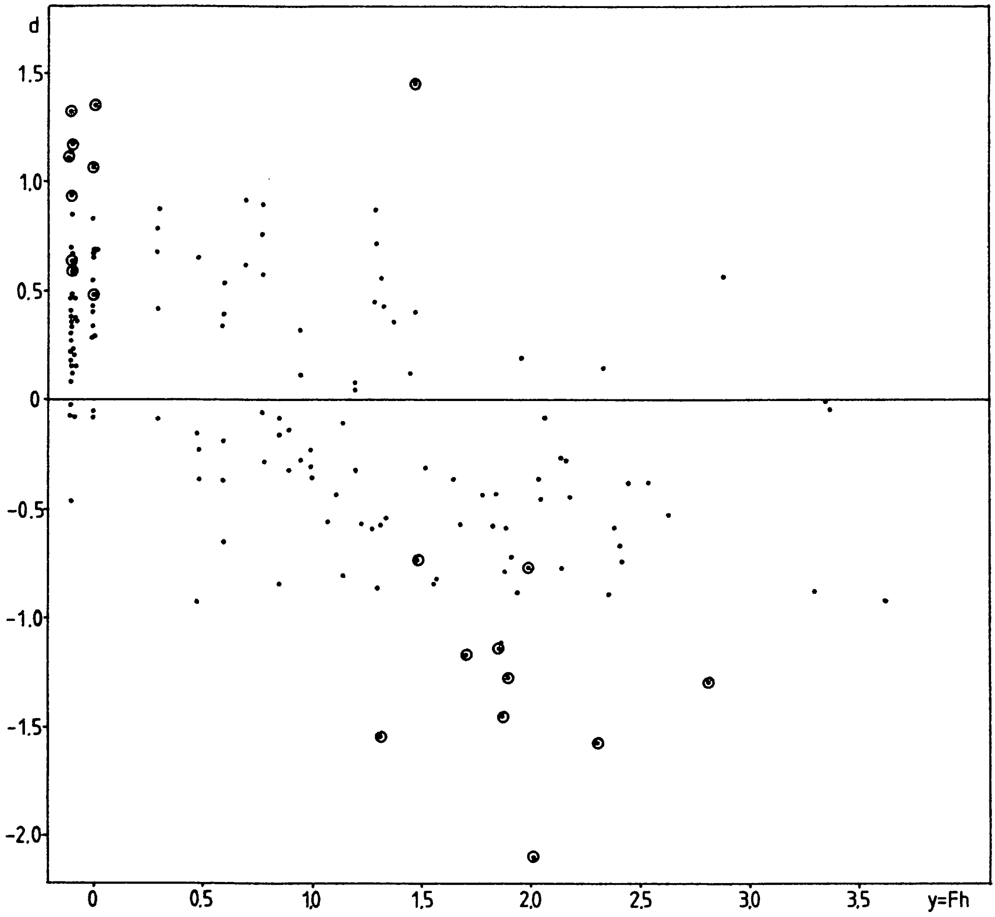

Fig. 5. Like Fig. 4, for the predicted variable $y = Fh$.

REFERENCES

Antoch, J., Bartkowiak, A., and Pękalska, J. 1986, *Report N-159*, Inst.of Computer Science, Wrocław University, March, 1986.

Antoch, J., Collomb, G., and Hassani, S. 1984, in *Proceedings of COMPSTAT 84*, Havranek *et al.* , eds. Physica Verlag Vienna, pp. 49–54.

Bartkowiak, A., and Jakimiec, M. 1986, in *Sol.-Terrest. Pred. Proc.*, P.A. Simon, G. Heckman and M. Shea, eds. (Boulder, USA), p. 285.

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

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

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

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. 1986, *Robust Statistics - The Approach Based on Influence Function*, (Wiley, New York).

Huber, P.J. 1981, *Robust Statistics*, (Wiley, New York).

Jakimiec, M., and Bartkowiak, A. 1989, *Acta Astr.*, **39**, 257.
Jakimiec, M., and Bartkowiak, A. 1990, *Acta Astr.*, **40**, 159, this issue.
Jakimiec, M., and Wanke-Jakubowska, M. 1988, *Acta Astr.*, **38**, 431.