WHAT IS ADAPTIVITY OF REGRESSION ANALYSIS INTENDED FOR?

J. Á. Víšek
ÚTIA ČSAV, Pod vodárenskou věží 4, 182 08 Praha 8

1. Abstract

Reasons for construction of adaptive statistical procedures are discussed. It is shown that increase of efficiency may be really important only exceptionally. Another, much more relevant reason may be selection of such model which gives a high hope to bring really consistent estimates (not estimators).

2. Introduction

It was in 1955 when C. Stein (1956) challenged statisticians with a question of possibility of efficient estimation (of parameters) when corresponding distribution is (completely) unknown. Solutions for location parameter problem under assumption of symmetry of underlying distribution were given by R. Beran (1978) and C. Stone (1975). For a general (theoretical) solution see Bickel (1982). Naturally such estimation needs to adapt estimating procedure to unknown underlying distribution. It brings some (technical) complication and evaluation of estimates are more time and space consuming. Hence there is a question whether such estimation has also other than theoretical sense. The last sentence shouldn't say that if solution of the problem of adaptive estimation will occur interesting only from theoretical point of view that it may be (or even should be) considered that it is too little to try to solve it. On the other hand there can appear another reason(s) important from practical point of view.

Since the question was formulated by C. Stein as the question of possibility to attain (asymptotically) efficiency the very first idea may be that adaptive estimation should (considerably) increase efficiency, i.e. that it elevates the amount of exploited information brought by data a lot.

In another words the first hope is that when employing adaptive procedure we may decrease losses we suffer when making use of procedures not fully appropriate for underlying (unknown) distribution.

The fact that our procedure is not fully fitting for underlying (type of) distribution may be due to it that we either "guessed" wrongly the kind of uncertainty (assuming e.g. normality of errors) or deliberately turned to some procedures with not so strict assumptions about underlying distribution and paid for it by a decrease of efficiency. As an example of such intently chosen but (possibly) less efficient procedure may serve a nonparametric or robust one. On the other hand since commonly known procedures of these types (as for instance median as a robust point estimator of location) really have (or may have) rather low efficiency it is unconsciously accepted that when we use any robust procedure we unavoidably loose a lot (of efficiency).

Although very high quality data usually contain no gross errors they still tend to have heavier tails than normal distribution. Let us mention that such excellent statisticians as K. Pearson (1902), Student (1927) and Jeffreys (1939) studied models which are the most approprimate to explain high quality data and they found that it needs longer-tailed distribution than the normal one.

K. Pearson collected series of data under highly uniform conditions and he found that the best approximation of their distribution function is by Student t with 5 - 9 degree of freedom. In recent time also P. J. Huber came to the same conclusion that suitable model for some kind of high-quality data can look like t_3.

In the light of this knowledge it may be interesting to recall that Fisher (1922) derived that asymptotic efficiency of mean under t_ν is $1 - 6/[\nu(\nu+1)]$ and that of variance is $1 - 12/[\nu(\nu+1)]$. Calculating corresponding values for t_9, t_5 and t_3 you obtain for mean 93 %, 80 % and 50 %, respectively, and for variance even 83 %, 40 % and 0 % (!) (this is an <u>asymptotic</u> value for $\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ when $n \to \infty$). So it seems at a first glance that adaptivity may improve the situation rather much. But on the other hand for mixture model $F(x) = (1-\varepsilon)\,\Phi(x) + \varepsilon\Phi(x/3)$ (Φ being standard normal distribution) we may find for ε running from 0 % to 10 % that the efficiency of mean decreases from 1 (i.e. 100 %) to 70 %, while that of median increases from 63,7 % approximately to 70 %. It means that the mean is uniformly better (for these contamination levels) than the median. Naturally this uniform majorization may be dreadfully broken by presence even one gross error. On the other hand 6 %-trimmed mean has efficiency not lower than 96 % over the whole interval 0 - 10 % of contamination. And moreover, presence of less than 6 % of gross errors doesn't cause catastrophe. Nevertheless if we are willing to pay more (money and time) for evaluation of estimate we may have even higher efficiency and higher safety again gross errors (let us mention as example some types of one or two steps M-estimators). For more details and another information see Hampel et al. (1986).

So the last example showed that application of an even simple robust estimator may yield such level of efficiency we will be satisfied with. Generally if we select from a family of robust procedures such one which is (at least approximately) optimal for given contamination level (or expected range of contamination levels) the loss of efficiency will be probably such that any other activity oriented on an increase of efficiency may occur as nonefficient and should be given up. (For the possibility of estimation of contamination level and following procedure selection see Víšek (1989), (1983)).

Therefore an increase of efficiency as a reason for adaptivity may be rather exceptional. Such increase may be considered as a sufficient reason for use of an adaptive procedure only in the case of very wide range of contamination levels or high uncertainty about type of distribution (of the bulk of data). Also in the case when we have at our disposal rather simple and quick adaptive procedure (and sufficiently large data) we may use it.

Nevertheless there is probably another reason for construction and application of some adaptive procedure which may be considered very important. It will be discussed in what follows. Let us restrict ourselves on point estimation.

One very important feature of statistical estimates is that they are "near" the "true" value. This may be expressed in mathematical model by unbiasedness. But restriction on the unbiased estimators is usually rather drastic if not impossible. Then we ask for a weaker form of the above mentioned requirement, namely consistency.

Consistency is sometimes viewed as something what is an automatically possessed property of any estimator (although for some kind of estimators it may be difficult to prove it). Hence we commonly appreciate much more asymptotic

normality (and maybe other characteristics of estimator as low gross-error-sensitivity or high breakdown point). But in fact the consistency is the only our guarantee that we are not too far from a "true model". Hence we should verify as far as possible very carefully those conditions under which consistency holds. It seems to be basic point of data analysis, point which may decide whether the whole analysis of data will be reasonable or not and the results reliable or not. On the other hand usually these conditions are to be fulfilled for unknown distribution and hence unverifiable (or verifiable only mediately, sometimes on a basis of vague knowledge having source in (physical) circumstances) . But let us consider linear regression analysis which the rest of paper will be devoted to. (Naturally, conclusion(s) will hold for location parameter - as a special case of regression-as well.)

For given data we may apply selected procedures of estimating regression coefficients and then having at hand corresponding residuals, an estimator of density may be applied which can give a hint whether or not conditions for consistency were fulfilled. What does the last sentence want to say? Let us consider classical least squares. After applying them on data one may verify whether the assumption of normality of residuals (which is basic for LS) has been (at least approximately) fulfilled. (One can test it formally by some test of fit but due to the fact that residuals are not mutually independent it is necessary carefully judge what the result of, let us say, χ^2 test means.)

Let us turn now our attention to one of (probably) basic condition of consistency, namely symmetry of residuals distribution. Some papers assume directly that the distribution of residuals is symmetric. Let us mention as an example Bickel (1975), Rouseeuw and Yohai (1984), Ruppert and Carroll (1980) or Yohai (1974) among others. May be that much more frequently the assumption of symmetry was made in papers devoted to location problem - see e.g. Jurečková (1983), Jurečková and Sen (1981). In the case of location parameter the assumption of symmetry brought not only a technical simplification but also improvement of heuristic behind the problem since in the case of symmetry the mean, the median and usually also the modus of density (if unimodal) coincide with the center of symmetry and hence it is (probably) out of any discussion what is to be estimated. For linear regression this fact seems to be transformed into the problem whether we want to estimate only slopes (then we may probably do without symmetry quite well) or we want to estimate simultaneously slopes and intercept. Then dispense with symmetry may cause at least technical troubles.

Moreover a lot of theoretical results simplify their form (or even are analytically expressable only) under assumption of symmetry. See e.g. Jurečková (1977) (and Huber (1969)), Remark after Corollary 3.1 which shows that M and R estimates have an asymptotically minimax property over the set of asymptotically unbiased estimates in the model of symmetric contamination. Another example may be found in Hampel et al (1986). Since location component of influence function and asymptotic variance (of corresponding estimator) under requirement of equivariance of estimator and symmetry of model distribution does not depend on the scale estimator used (as preliminary estimator of scale for instance to rescale data before using a standard M-estimator), the estimator is optimal independently of the preliminary estimation. Similarly the change-of-efficacy function is given only under assumption of symmetry.

As a last example let us mention the estimation of regression coefficients by means of regression quantiles (introduced by Koenker and Bassett (1978)).

Although this technic does not suppose any symmetry and even the number which is cut off may be different for upper and lower quantiles (see Ruppert and Carroll (1980)), it may be profitable to assume symmetry of residuals since otherwise the bias of estimator of intercept is nonzero. It seems natural that one would expect that if you find median quantile (i.e. quantile for $\alpha = \frac{1}{2}$) that it will estimate the "true" values of coefficients consistently. But as Jurečková (1984) proved the estimate of intercept converges in probability to $F^{-1}(\frac{1}{2})$ and it means that only for F symmetric the estimate of intercept is consistent.

So it seems that there are some more or less important reasons to assume symmetry (of residuals). But that all what was said up to now should not be understood so that symmetry must be assumed. Without any doubt there are sets of real data which need to be processed by method not assuming symmetry at all or at least the interpretation of the results of analysis using methods based on requirement of symmetry have to be accommodated to it. (See also discussion about a fiction - unfortunately largely spread among statisticians - that symmetry is basic assumption of any robust procedure; in Hampel et al. (1986) point 10. in chapter 8.2a.)

On the other hand if we have good reasons for assumption of symmetry it in fact may improve our data processing very much. In some sense it brings so much information that our results may behave as if we would have two times more data (see Eeden et al. (1985)).

Moreover it is sometimes useful to distinguish between symmetry of model for bulk of data (for the prevailing part of residuals) and symmetry or asymmetry of contamination.

Let us stop for a while and summarize that we have done up to now. We have collected some reasons which may support the idea that symmetry of model distribution may be very useful. But there may be an objection! The "true" model may have an asymmetric distribution of residuals! But what is true model? When we solve location or scale problem the answer seems to be quite simple and an idea about model may be created estimating the density. But for the problem of estimation of regression model any answer will not be so convincing. We are usually in a situation when we try to explain (structure of) data by some regression model and we may often select even from more mutually competing models. The physical circumstances which point out only one (type of) "true" model are probably seldom. In the case of choosing from a set of competitors the above given reasons may argue for preference of model with (approximately) symmetrically distributed residuals. It is not difficult to invent a statistics measuring level of symmetry.

In the next chapter one procedure which looks for (the "best") model (with symmetric residuals) is described. Another one may be then used to increase efficiency. It will be proposed, too.

3. Notation and Preliminaries

Let us consider a linear regression model

$$Y = X\beta^{O} + e$$

where $Y = (Y_1,\ldots,Y_n)'$ is a response variable, $X = \{x_{ij}\}_{i=1,\,j=1}^{n\quad p}$ is a design matrix (in the case that intercept is supposed to be included in model we assume $x_{i1} = 1$ for $i = 1,\ldots,n$), $\beta^{O} = (\beta_1^{O},\ldots,\beta_p^{O})'$ vector of regression coef-

ficients, and $e = (e_1,\dots,e_n)'$ is a vector of i.i.d. (according to some distribution G) random variables. (We assume that $p \geq 2$.) G is assumed to allow absolutely continuous density g being symmetric. Moreover we suppose that Fisher information is finite and denote it by I(g). Let R denote real line and N the set of all positive integers.

Both estimators of regression coefficients will be based on the kernel estimator of density of residuals. So we need a necessary notation for it. Let for any $i = 1,\dots,n$ X_i denotes i-th row of design matrix and for any $\beta \in R^p$ let $e_i(\beta) = Y_i - X_i\beta$ be i-th residual. Let w be a symmetric and everywhere positive kernel and $\{c_n\}_{n=1}^{\infty}$ a sequence of positive numbers converging to zero. Denote for any $y \in R$

$$g_n(y,Y,\beta) = \frac{1}{nc_n} \sum_{i=1}^{n} w\left(c_n^{-1}(y-[Y_i-X_i\beta])\right)$$

the kernel estimator of density of residuals. For the kernel we shall assume that it is three times absolutely continuous and that there exist constants $K_1,\dots,K_5$ such that

$$\sup_{y\in R} w(y) < K_1 , \qquad \sup_{y\in R} \frac{|w'(y)|}{w(y)} < K_2 ,$$

$$\sup_{y\in R} \frac{|w''(y)|}{w(y)} < K_3, \qquad \sup_{y\in R} \frac{|w''(y)|}{w(y)} < K_4$$

and for any $n \in N$

$$\max_{\substack{i=1,\dots,n\\ j=1,\dots,p}} |x_{ij}| < K_5 .$$

4. Estimator based on Hellinger distance

For a sequence $\{a_n\}_{n=1}^{\infty}$ of positive numbers denote by $b_n(y)$ a symmetric differentiable function such that for all $y \in R$ $0 \leq b_n(y) \leq 1$ and

$$b_n(y) = \begin{cases} 1 & |y| \leq a_n , \\ 0 & |y| > a_n + c_n^4 . \end{cases}$$

Definition 1. For any $Y \in R^n$ put

$$\hat{\beta}^n(Y) = \operatorname*{argmax}_{\beta\in R^p} \int g_n^{\frac{1}{2}}(y,Y,\beta)\, g_n^{\frac{1}{2}}(-y,Y,\beta) b_n(y)dy .$$

If there is not any such point then put $\hat{\beta}^n(Y)$ equal to any $\tilde{\beta} \in R^p$ such that

$$\int g_n^{\frac{1}{2}}(y,Y,\tilde{\beta})\, g_n^{\frac{1}{2}}(-y,Y,\tilde{\beta})dy > \sup_{\beta\in R^p} \int g_n^{\frac{1}{2}}(y,Y,\beta)\, g_n^{\frac{1}{2}}(-y,Y,\beta)\, b_n(y)dy - \frac{1}{n} .$$

Denote for any a, b $\in R$ by $\mathcal{E}_p(a,\beta^0) = \left\{\beta \in R^p:\ \|\beta-\beta^0\| > a\right\}$ and by

$$\mathcal{E}_p(a,b,\beta^0) = \mathcal{E}_p(a,\beta^0) \cap \mathcal{E}_p^c(b,\beta^0) .$$

Condition A. For any $\delta > 0$ there is $\Delta \in (0,1)$ and $K_\Delta \in R$ such that

i)

$$\limsup_{n\to\infty} \sup_{\mathcal{E}_p(\delta,K_\Delta,\beta^0)} \int E^{\frac{1}{2}} g_n(y,Y,\beta)\, E^{\frac{1}{2}} g_n(-y,Y,\beta)\, b_n(y)dy < \Delta$$

and

ii)

$$\limsup_{n\to\infty} \sup_{\mathcal{E}_p(K_\Delta,\beta^0)} \int g_n^{\frac{1}{2}}(y,Y,\beta)\, g_n^{\frac{1}{2}}(-y,Y,\beta)\, b_n(y)dy < \Delta \text{ in probability.}$$

Remark. Condition ii) is meant so that the left hand side converges in probability to a random variable less then Δ.

Theorem 1. Let Condition A be fulfilled and

$$\lim_{n\to\infty} c_n = 0, \quad \lim_{n\to\infty} nc_n = \infty \quad \text{and} \quad \lim_{n\to\infty} nc_n^{4p}a_n^{-2p} = \infty .$$

Then $\hat{\beta}^n(Y)$ is (weakly) consistent estimator of β^0.

Theorem 2. Let there is $M \in R$ such that

$$\sup_{y\in R} |g'(y)| < M$$

and

$$\limsup_{n\to\infty} g(a_n) / c_n = 0 .$$

Further let

$$\int t\, w^2(t)dt < \infty$$

and Condition A be fulfilled. Then

$$n^{-\frac{1}{2}}\, I(g) \sum_{\ell=1}^{p} \left(\hat{\beta}_\ell^n - \beta_\ell^0\right) \sum_{i=1}^{n} x_{i\ell} = n^{-\frac{1}{2}} \sum_{i=1}^{n} g'\left(Y_i - X_i\beta^0\right) g^{-1}\left(Y_i - X_i\beta^0\right) + \sigma_p(1).$$

5. Maximum-likelihood-like estimator

In what follows let $\tilde{\beta}^n$ denote a preliminary estimator of regression coefficients and denote by $\tilde{e}_i$ residuals $e_i(\tilde{\beta}^n)$. For a sequence $(a_n)_{n=1}^{\infty}$ of positive number, $a_n \nearrow \infty$ define

$$b_n(y) = \begin{cases} 1 & |y| \le a_n , \\ 0 & |y| > a_n . \end{cases}$$

Condition B. Let

$$\lim_{n\to\infty} n^{\delta}c_n^4 = \infty \quad \text{and} \quad \lim_{n\to\infty} n^{4\delta-1}c_n^6 = \infty .$$

Moreover let

$$n^{2\delta}E\|\tilde{\beta}^n - \beta^0\|^2 = \sigma(1)$$

and preliminary estimator is assumed to be such that for any $j = 1,\dots,n$; $i = 1,\dots,n$, $t \in R$, $s \in R$

$$P_g\left(\tilde{e}_i - E(\tilde{e}_i \mid e_j=t) < s\right) = P_g\left(\tilde{e}_i - E(\tilde{e}_i \mid e_j=t) > -s\right) .$$

Condition C. Let for any $a \in R$

$$\lim_{n\to\infty} \sup_{|b|<a} n^{-\frac{1}{2}}\, c_n^{-2} \int w^{-1}\left(c_n^{-1}(z+b-t)\right) g(t)\, g(z)\, dtdz = 0 .$$

Further let there exist ν, D ($\nu > 0$, $D > 0$) such that for any z_1, $z_2 \in R$, $|z_1-z_2| < \nu$ we have $w(z_1)/w(z_2) < D$. Let

$$\lim_{n\to\infty} \frac{1}{n} X'X = Q$$

where Q is a regular matrix.

Definition 2. For any sequence $(d_n)_{n=1}^{\infty}$ denote by

$$G\left((d_n)_{n=1}^{\infty}\right) = \Big\{h;\ h \text{ is a density and for any } n \in N$$

$$P_h\left\{\max\left\{\sup_{y\in R} |g_n(y,Y,\beta^0) - h(y)|,\ \sup_{y\in R} |g_n(y,Y,\beta^0) - E_h g_n(y,Y,\beta^0)|\right\} > \frac{1}{2}\, d_n\right\} < d_n\Big\} .$$

Definition 3. Under $\hat{\tilde{\beta}}^n$ we shall understand a point (or points) for R^p for which

$$\prod_{j=1}^{n} g_n\left(e_j(\beta), Y, \tilde{\beta}^n\right) b_n(\tilde{e}_j) = \max ! .$$

Again if such point does not exist let $\hat{\beta}^n$ be any point $\tilde{\beta}$ for which

$$\prod_{j=1}^{n} g_n\left(e_j(\tilde{\beta}), Y, \tilde{\beta}^n\right) b_n(\tilde{e}_j) > \sup_{\beta \in R^p} \prod_{j=1}^{n} g_n\left(e_j(\beta), Y, \tilde{\beta}^n\right) b_n(\tilde{e}_j) - \frac{1}{n} .$$

Condition D. Let

$$\lim_{n\to\infty} \frac{d_n}{c_n} = \infty .$$

Further let us assume that there is a constant K_6 such that

$$P_g\left(\|\hat{\beta}^n\| > K_6\right) \longrightarrow_{n\to\infty} 0 .$$

Let $g \in G\left((d_n)_{n=1}^{\infty}\right)$ and the above given sequence $(a_n)_{n=1}^{\infty}$ be chosen so (and g be of such type) that starting from some $n_0 \in N$ we have for any $n \geq n_0$

$$\left(-2a_n, 2a_n\right) \subset \left\{y \in R : g(y) > d_n^{\frac{1}{2}}\right\} .$$

Finally let

$$\lim_{n\to\infty} n\, c_n^6\, a_n^{-2} = \infty .$$

Condition E. Let the density g and the kernel w be such that for any $t > 0$

$$\int \left[g'(t-y) - g'(t)\right] w(y)\, dy > 0 .$$

Remark. Although the above given **Conditions B - E** may seem at the first glance rather restrictive they can be fulfilled for rather broad class of distributions. The details were described in Víšek (1990b).

Theorem 3. Let **Conditions B - E** holds. Then $\hat{\beta}^n(Y)$ is (weakly) consistent estimator of β^0 and the following representation takes place for any $k = 1,\ldots,p$

$$n^{-\frac{1}{2}}\left(\hat{\beta}^n - \beta^0\right)X'X_k = n^{-\frac{1}{2}} I^{-1}(g) \sum_{j=1}^{n} x_{jk} \frac{c_n^{-1}\int w'\left(c_n^{-1}(e_j - z)\right) g(z)dz}{\int w\left(c_n^{-1}(e_j - z)\right) g(z)dz} + o_p(1) .$$

Corollary.

$$\mathcal{L}\left(n^{-\frac{1}{2}}\left(\hat{\beta}^n - \beta^0\right)X'X\right) \longrightarrow_{n\to\infty} N\left(0,\ Q \cdot I^{-1}(g)\right) .$$

6. Discussion

It is clear from THEOREM 2 that the estimator based on Hellinger distance selects the model having - as far as possible - approximately symmetric distribution of residuals for the bulk of data. Since one of important assumption the proof of consistency was based on is symmetry of errors this method is consistent with its assumptions or at least it endeavours to be. On the other hand the second method - applying maximum likelihood principle on preliminary estimated distribution of residuals - has advantage that resulting estimates are asymptotically efficient. Both methods requires to find the estimates as point (or points) at which some functionals reach their maxima. When implementing these procedures one easy finds that it is possible to find such point only approximately by means of some iterative process starting from some

preliminary "guessed" values of coefficients. Experiences have shown that it may be useful to proceed as follows. At first we should find some (highly) robust estimate of regression coefficients e.g. LMS (Least Median of Squares) or LTS (Least Trimmed of Squares) estimate and starting from it to find (e.g. by means of steepest direction method) approximate solution of the first method. Then we may apply the second method which usually changes value of estimates only slightly (it is easy to see that for $\beta = \tilde{\beta}^n$ one obtains zero as a value of functional used for maximum-likelihood-like estimation and hence a solution are usually nearby).

Numerical illustration

As an numerical illustration we shall present two examples. The first one will consider an artificial data, the second then well known data - Stackloss data.

Let us start with the first example. To guarantee symmetry of "errors" the numbers were generated as follows. We have assumed regression model y = x + e with e distributed according to N(0,1.3). We have successively generated numbers uniformly distributed from [-4,4] (as values of x_i's) and normally distributed numbers (N(0,1.3)) as values of e_i's for i = 1,...,27. Then we have put $y_{2i} = x_i - |e_i|$. Further we have generated numbers ξ_i (uniformly on [0,2]) for i = 1,...,27 and put $y_{2i-1} = x_i - \text{sign}(x_i) \cdot \xi_i + |e_i|$. Finally we have constructed two leverage points in a similar way. For them we have changed sign of response variable. We have obtained

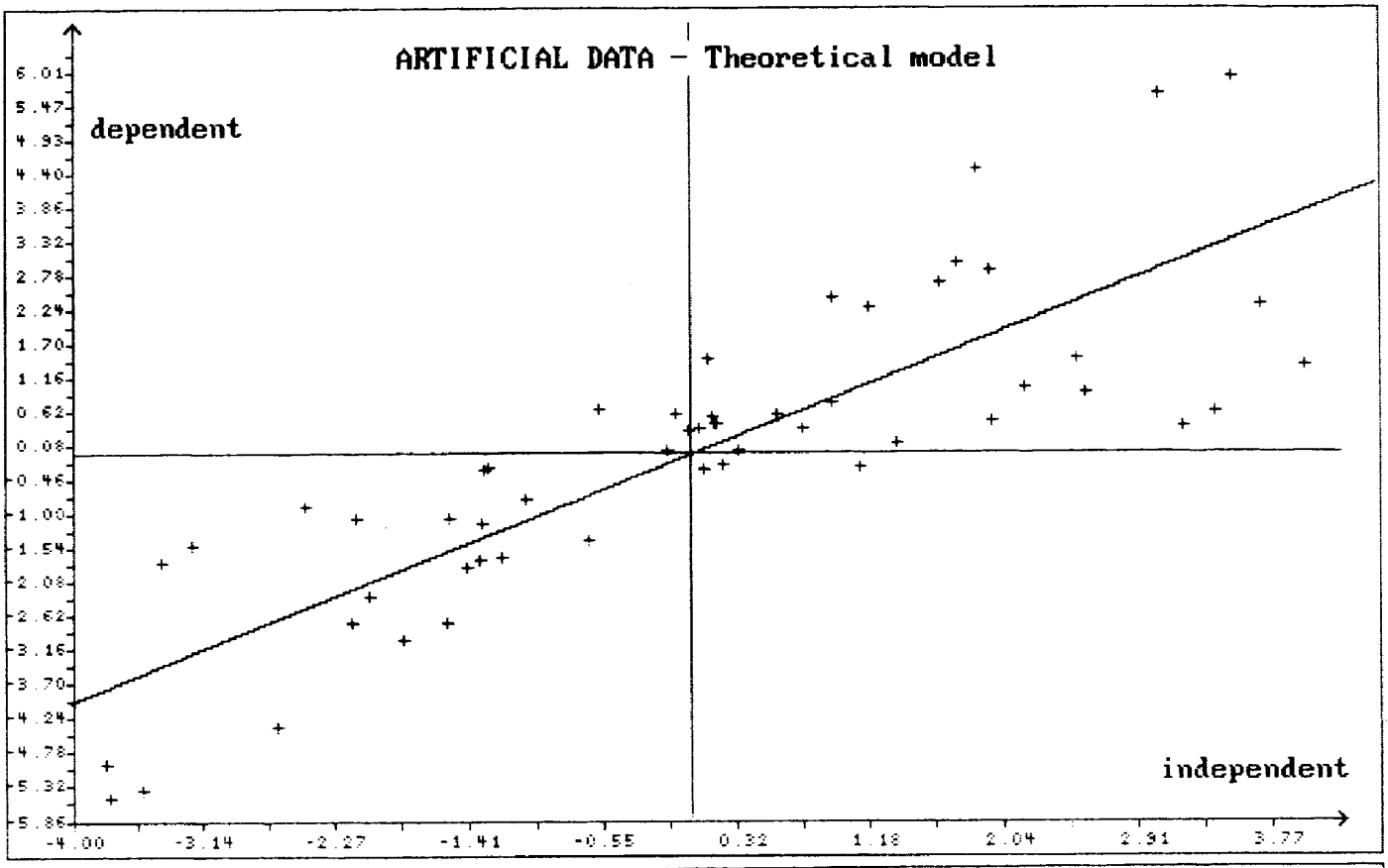
ARTIFICIAL DATA (y=x+e; x..U(-4,4); e..N(0,1.3); 2 leverage points)

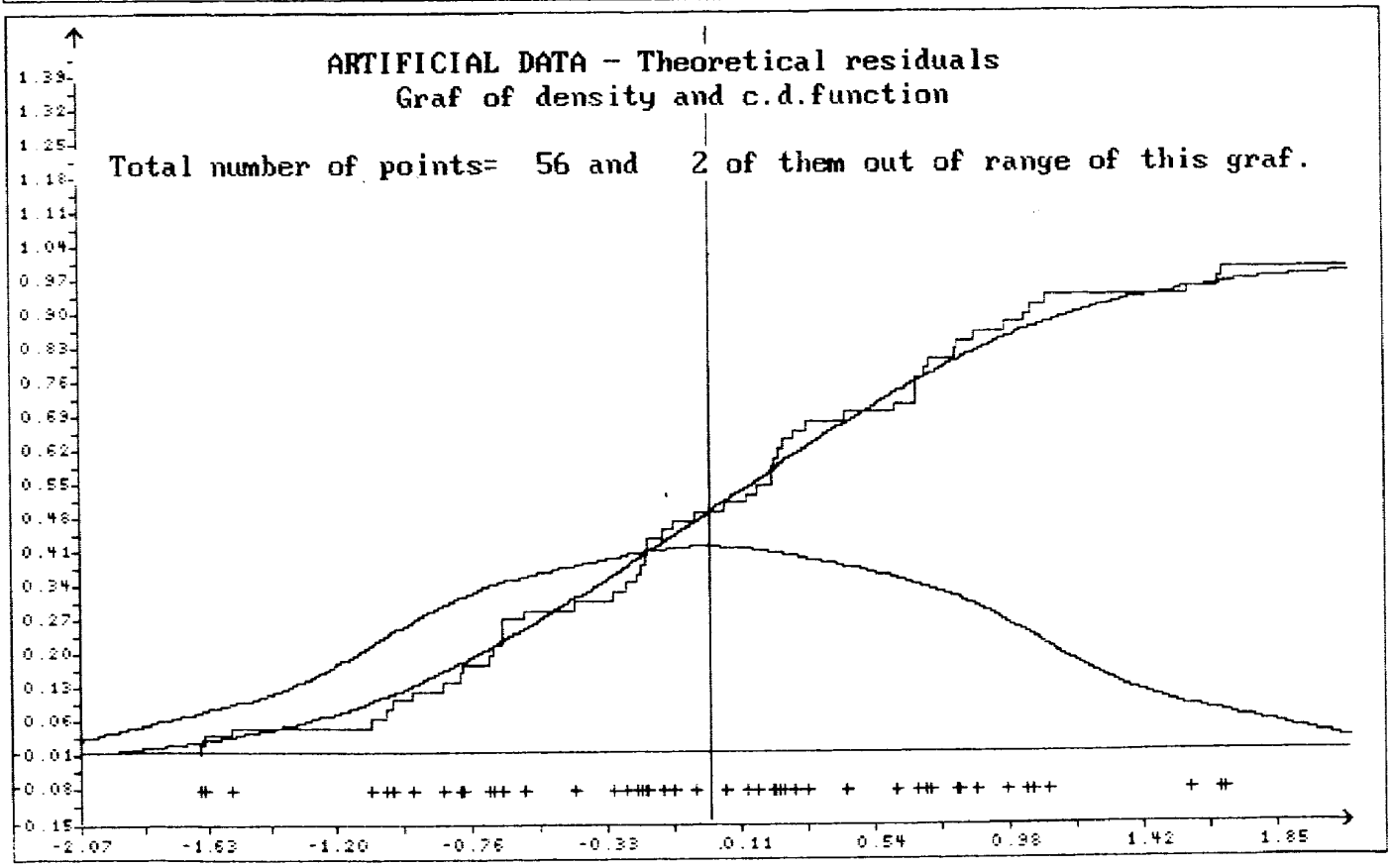
case	x	y	case	x	y
1	-3.758713	-5.450430	29	3.018787	5.729336
2	-3.419444	-1.727727	30	2.481405	1.506175
3	-1.443068	-1.807758	31	1.925544	2.900774
4	-1.069856	-0.705167	32	0.207729	-0.176134
5	-1.215064	-1.653717	33	-0.017527	0.366336
6	0.147676	0.586328	34	1.335698	0.173457
7	-2.664302	-4.326543	35	1.153117	2.315358
8	-2.481722	-0.819481	36	3.666636	2.355866
9	-3.780872	-4.921943	37	1.724177	3.034948
10	-2.152639	-1.011568	38	9.912632	-11.469360
11	-0.665691	-1.383678	39	1.099867	-0.204246
12	-0.089969	0.628017	40	-0.592642	0.711471
13	-1.564029	-2.658770	41	3.949518	1.425105
14	-1.334501	-0.239761	42	3.484233	6.008646
15	-1.840087	-2.936874	43	0.912645	0.830367
16	-1.295033	-0.198246	44	0.562032	0.644310
17	-2.173251	-2.683591	45	1.938431	0.532863
18	-1.549990	-1.039650	46	0.117101	1.522670
19	-1.357830	-1.690449	47	2.159913	1.063126
20	0.159062	0.491681	48	1.614859	2.711646
21	-3.551338	-5.326357	49	3.373824	0.688264
22	-3.221335	-1.446316	50	1.842544	4.528104
23	11.547648	-9.990920	51	0.303299	0.055823
24	-2.058207	-2.255504	52	-1.344843	-1.097367
25	-0.144130	0.053167	53	0.717807	0.389267
26	2.547648	0.990920	54	0.167538	0.496077
27	0.912632	2.469360	55	0.080927	-0.255941
28	3.170015	0.459466	56	0.055200	0.392068

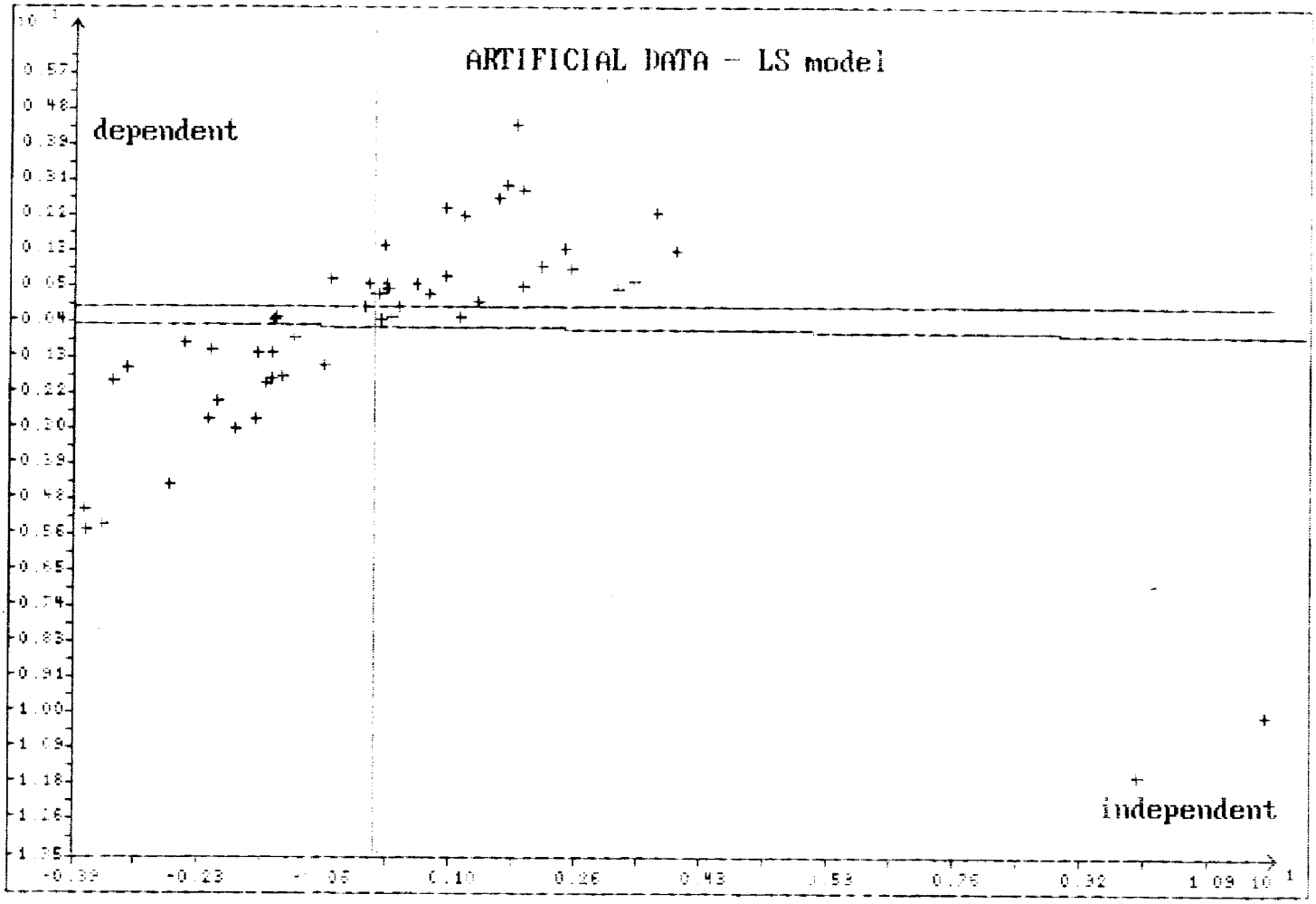
The LS method, GM-estimator (of gnostical type - see Novovičová (1990)), LMS estimator and adaptive estimator of Hellinger type were used. The result have been as follows

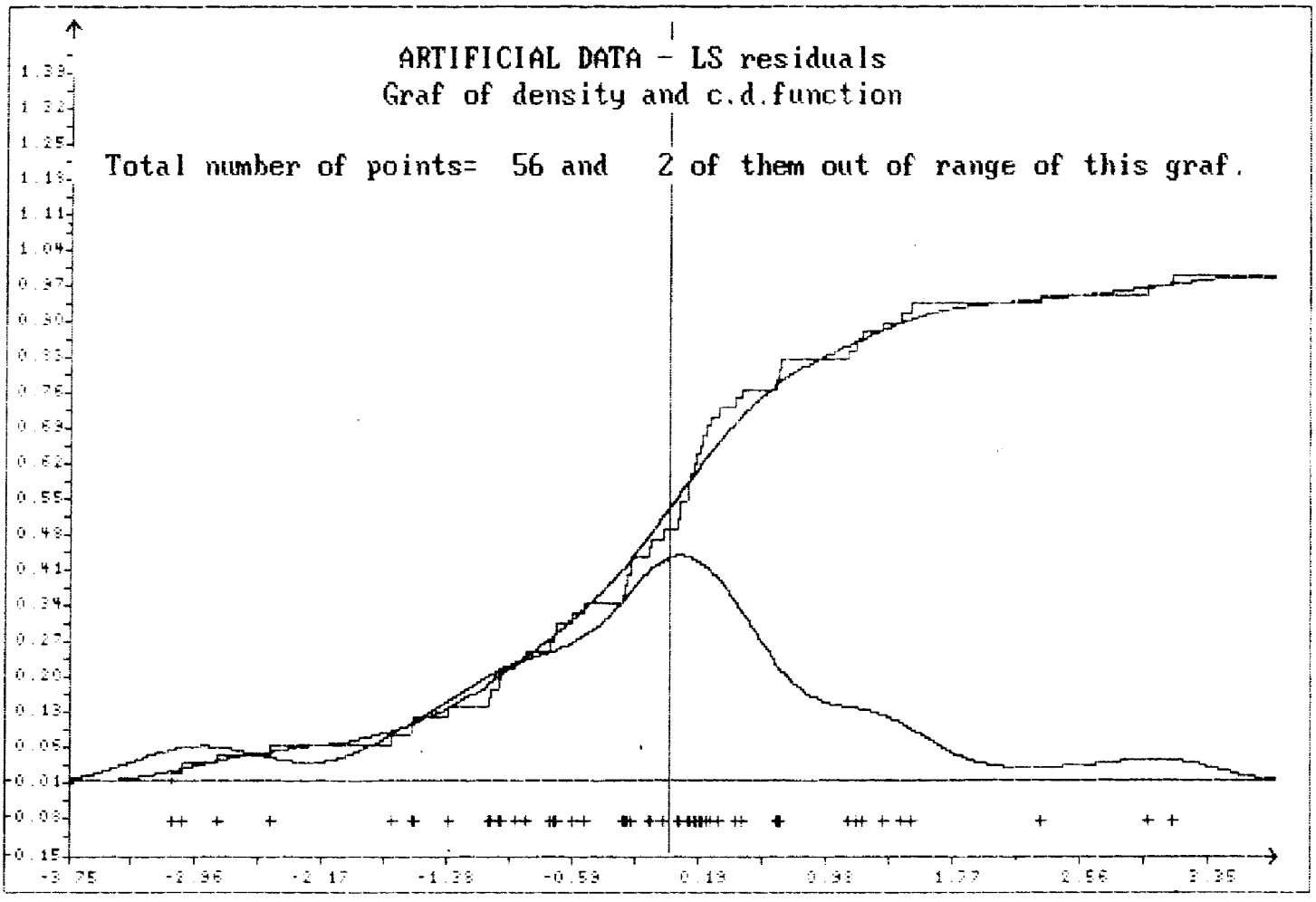
type of estimator	LS	GM	LMS	adaptive
intercept	-0.375	0.317	0.056	0.000
slope	-0.019	1.552	0.466	1.000

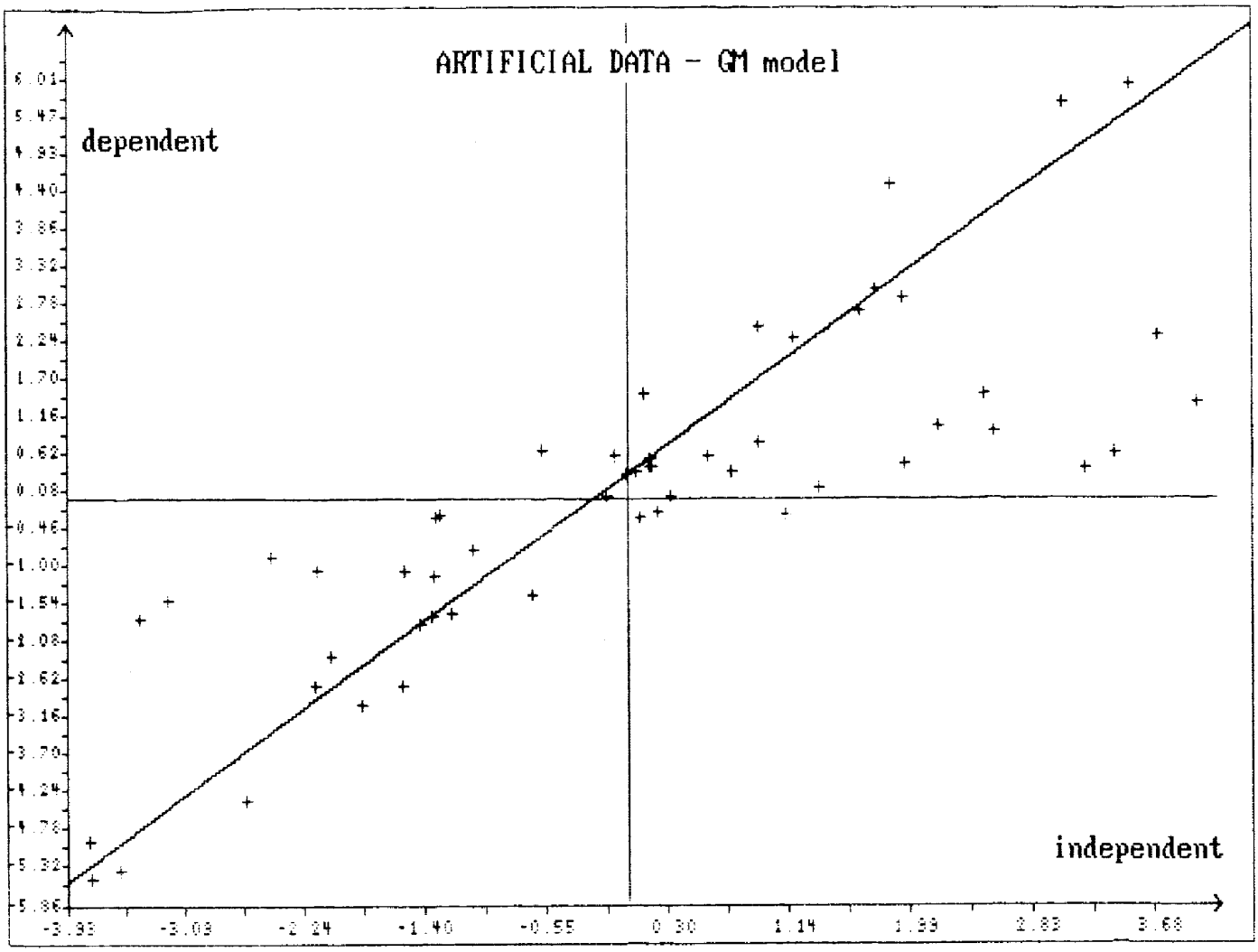
The following pictures should give an idea about estimated models and densities of residuals.

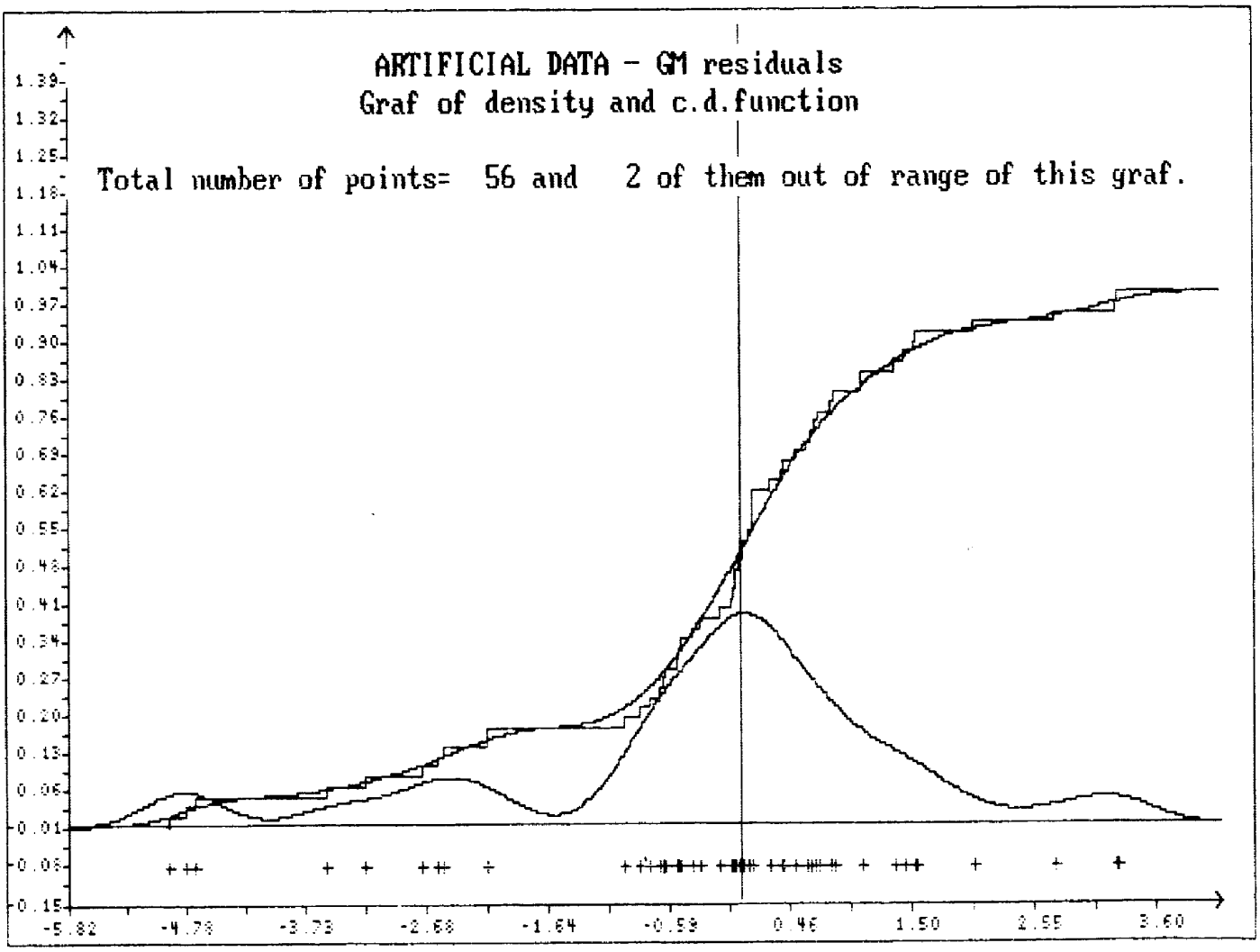


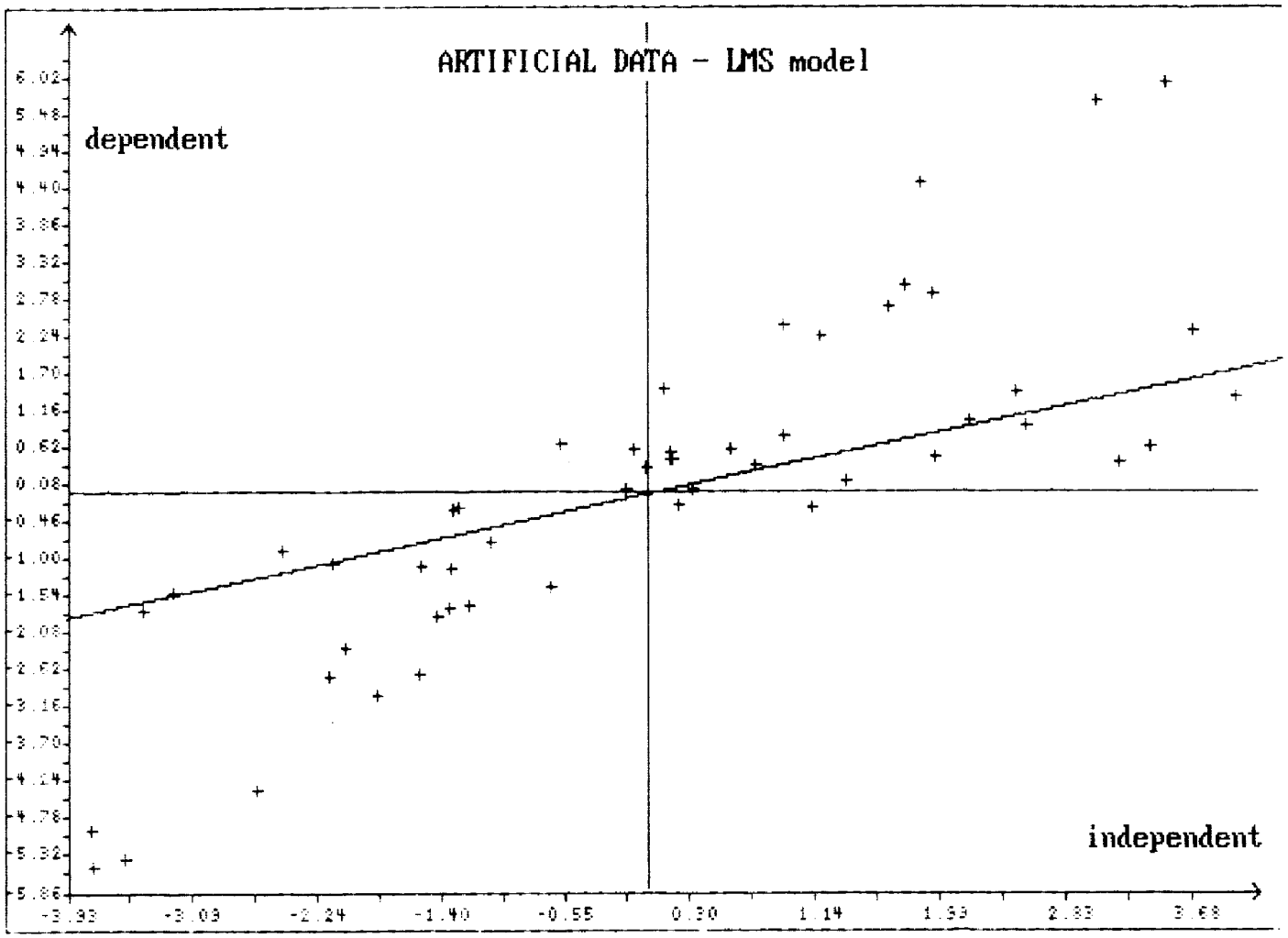


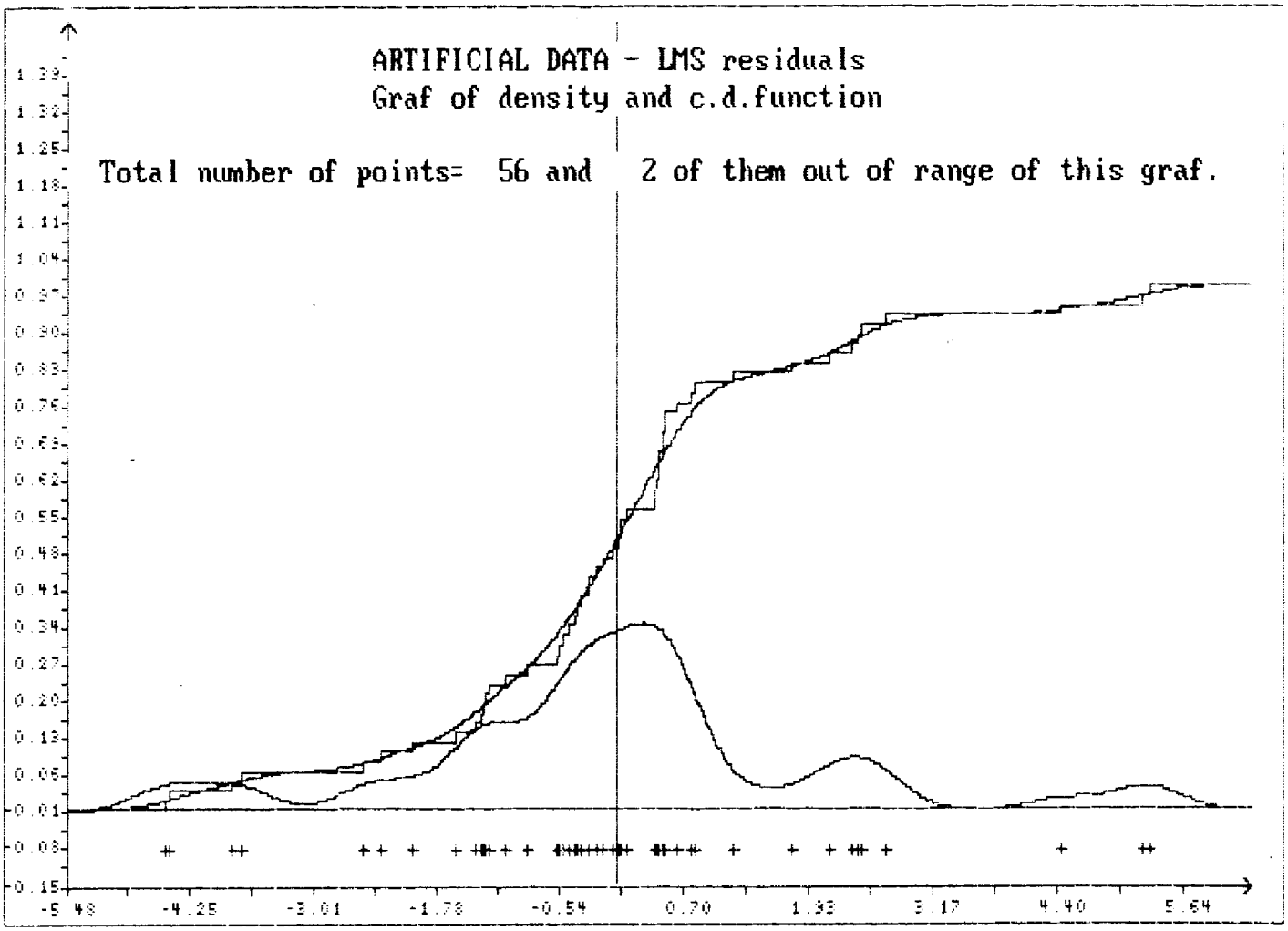

ARTIFICIAL DATA - LS model
dependent
independent

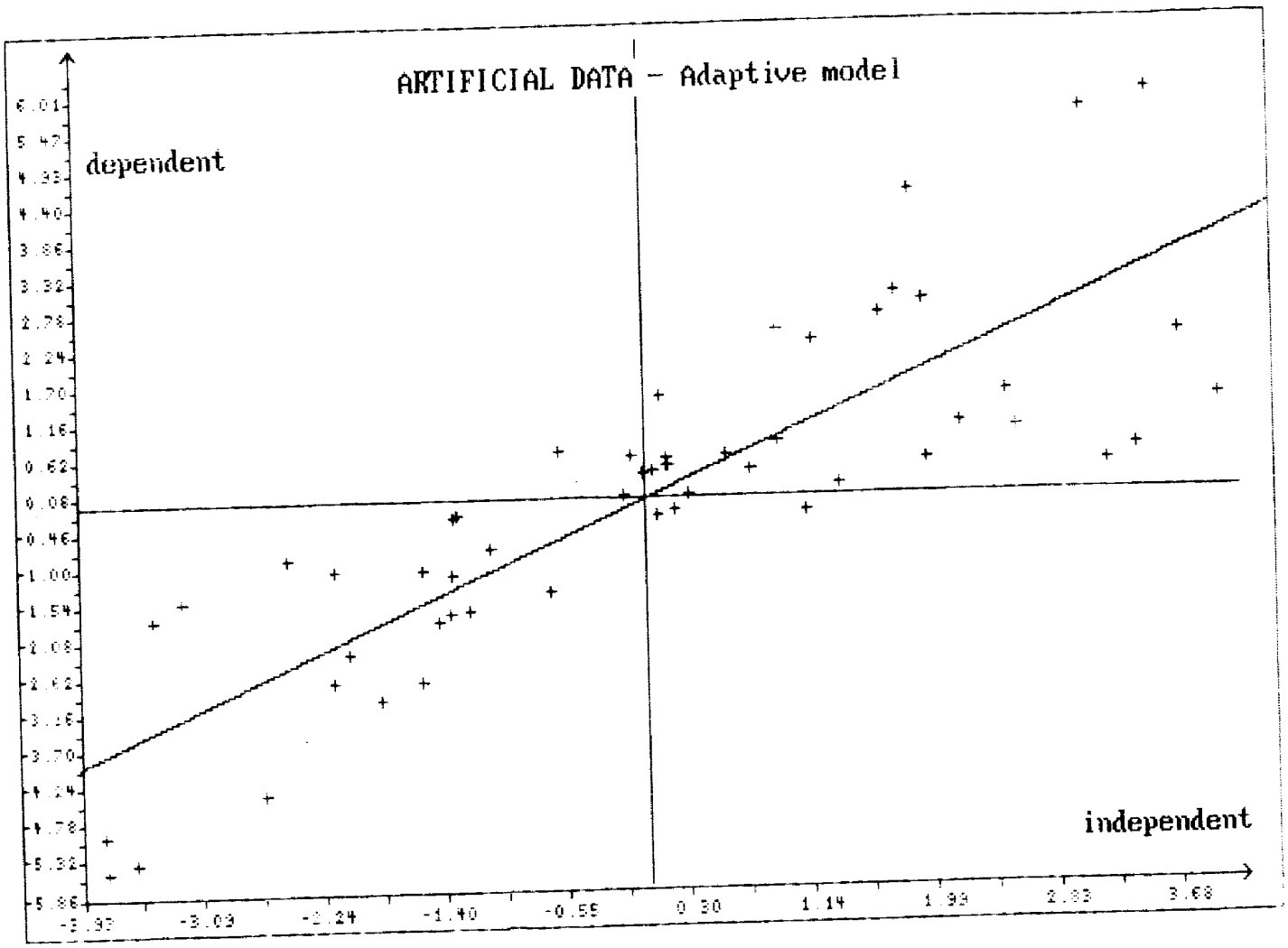

ARTIFICIAL DATA - LS residuals
Graf of density and c.d.function
Total number of points= 56 and 2 of them out of range of this graf.

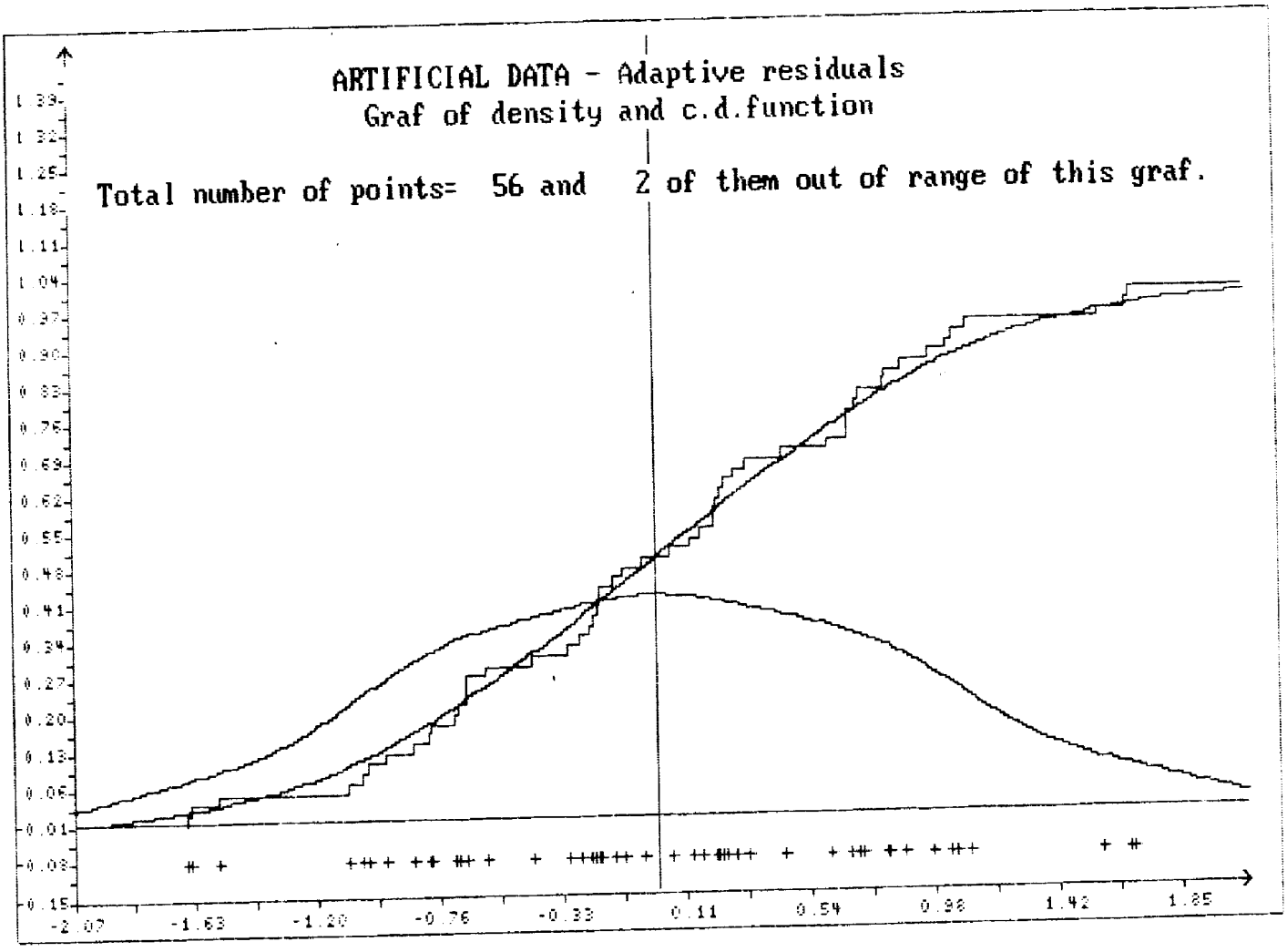

ARTIFICIAL DATA - GM model
dependent
independent
6.01
5.47
4.93
4.40
3.86
3.32
2.78
2.24
1.70
1.16
0.62
0.08
-0.46
-1.00
-1.54
-2.08
-2.62
-3.16
-3.70
-4.24
-4.76
-5.32
-5.86
-3.93
-3.09
-2.24
-1.40
-0.55
0.30
1.14
1.99
2.83
3.68


ARTIFICIAL DATA - GM residuals
Graf of density and c.d.function
Total number of points= 56 and 2 of them out of range of this graf.
1.39
1.32
1.25
1.18
1.11
1.04
0.97
0.90
0.83
0.76
0.69
0.62
0.55
0.48
0.41
0.34
0.27
0.20
0.13
0.06
-0.01
-0.08
-0.15
-5.82
-4.73
-3.73
-2.68
-1.64
-0.59
0.46
1.50
2.55
3.60


ARTIFICIAL DATA - LMS model
dependent
independent


ARTIFICIAL DATA - LMS residuals
Graf of density and c.d.function
Total number of points= 56 and 2 of them out of range of this graf.


ARTIFICIAL DATA - Adaptive model
dependent
independent

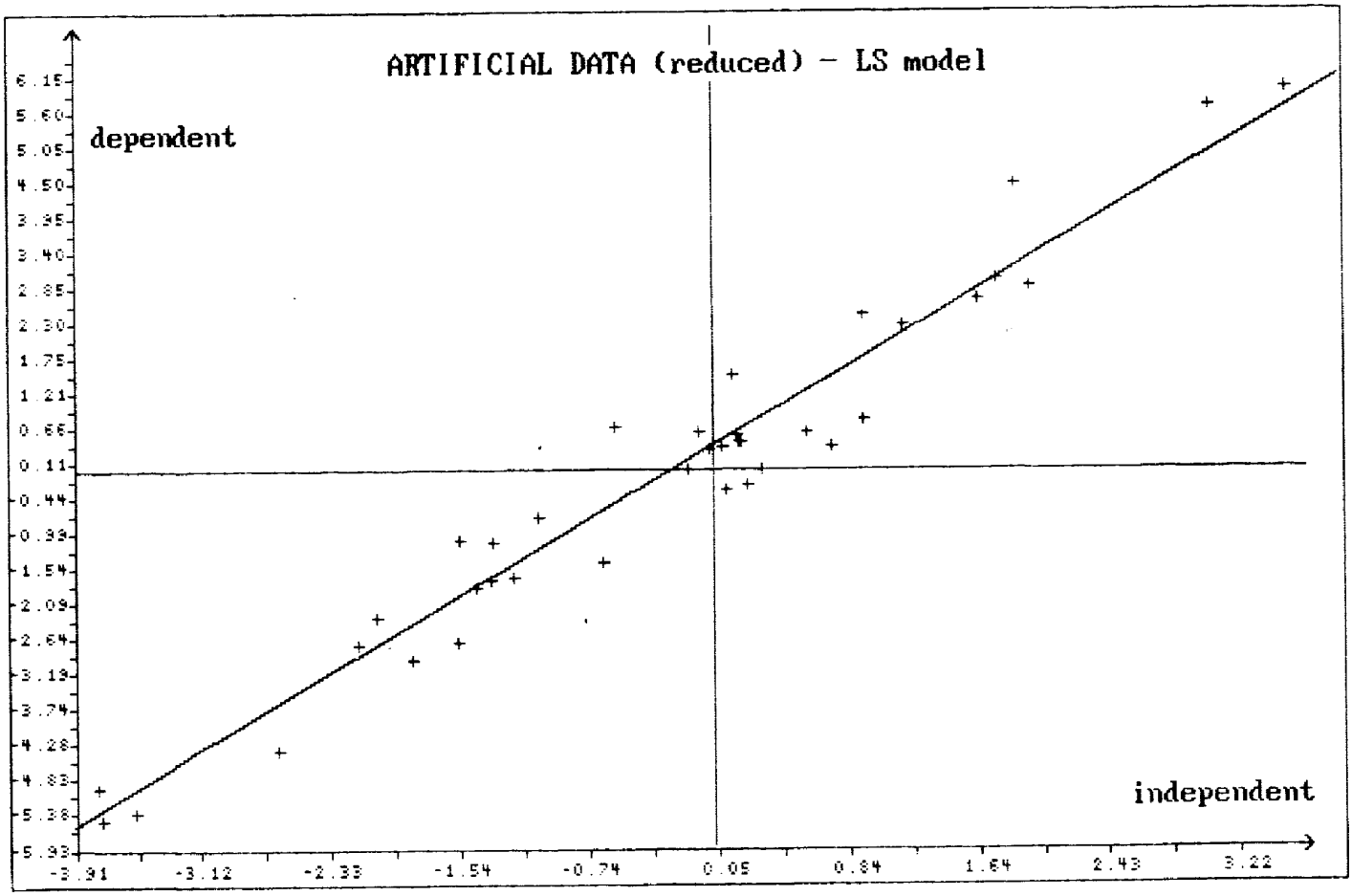

ARTIFICIAL DATA - Adaptive residuals
Graf of density and c.d.function
Total number of points= 56 and 2 of them out of range of this graf.

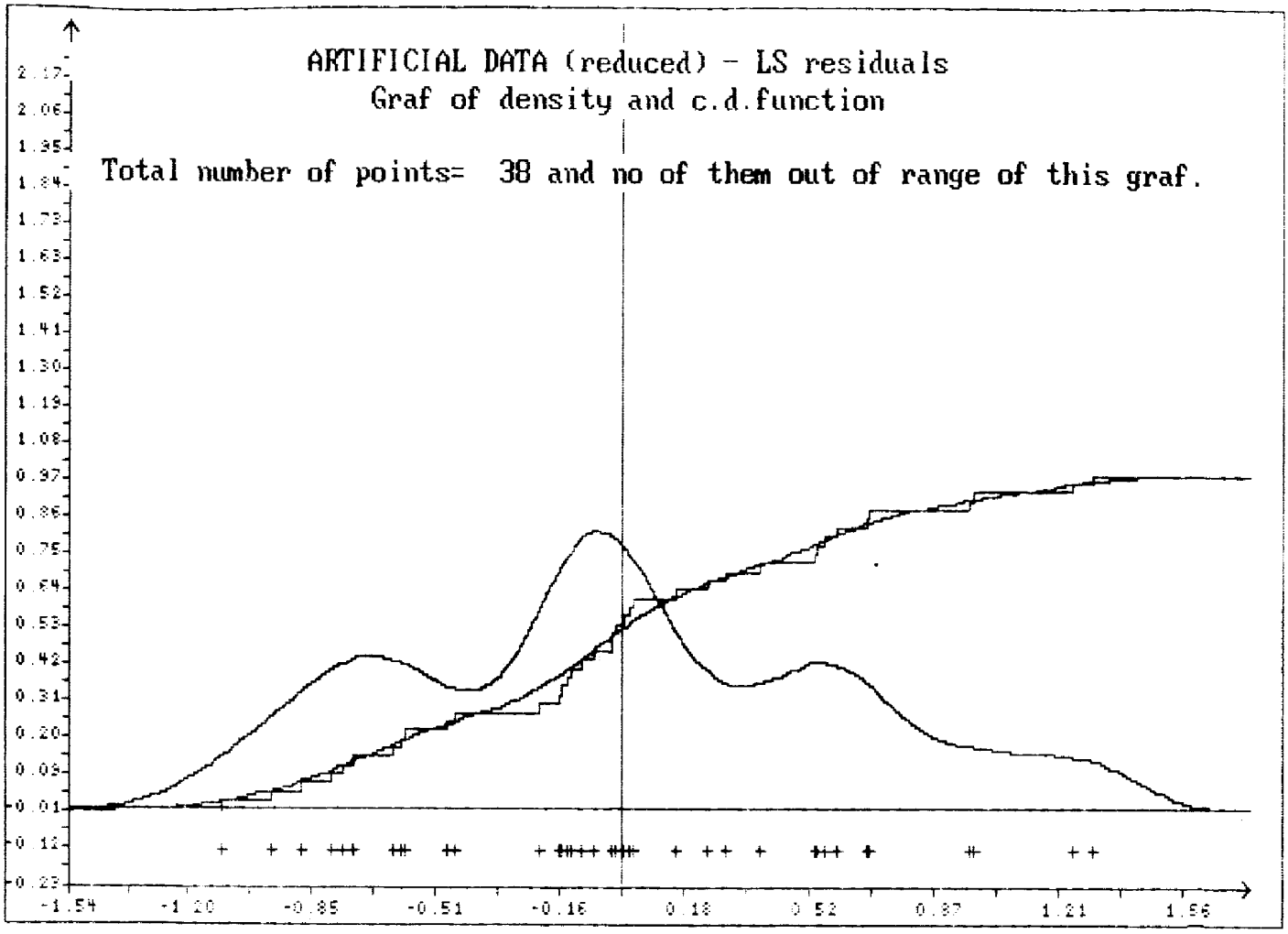
Since it seems that data could be viewed as a mixture of two groups one possible group was selected (according to GM-estimator) and again the above mentioned procedures were applied. The data and results (again with corresponding pictures) hawe been in this case as follows.

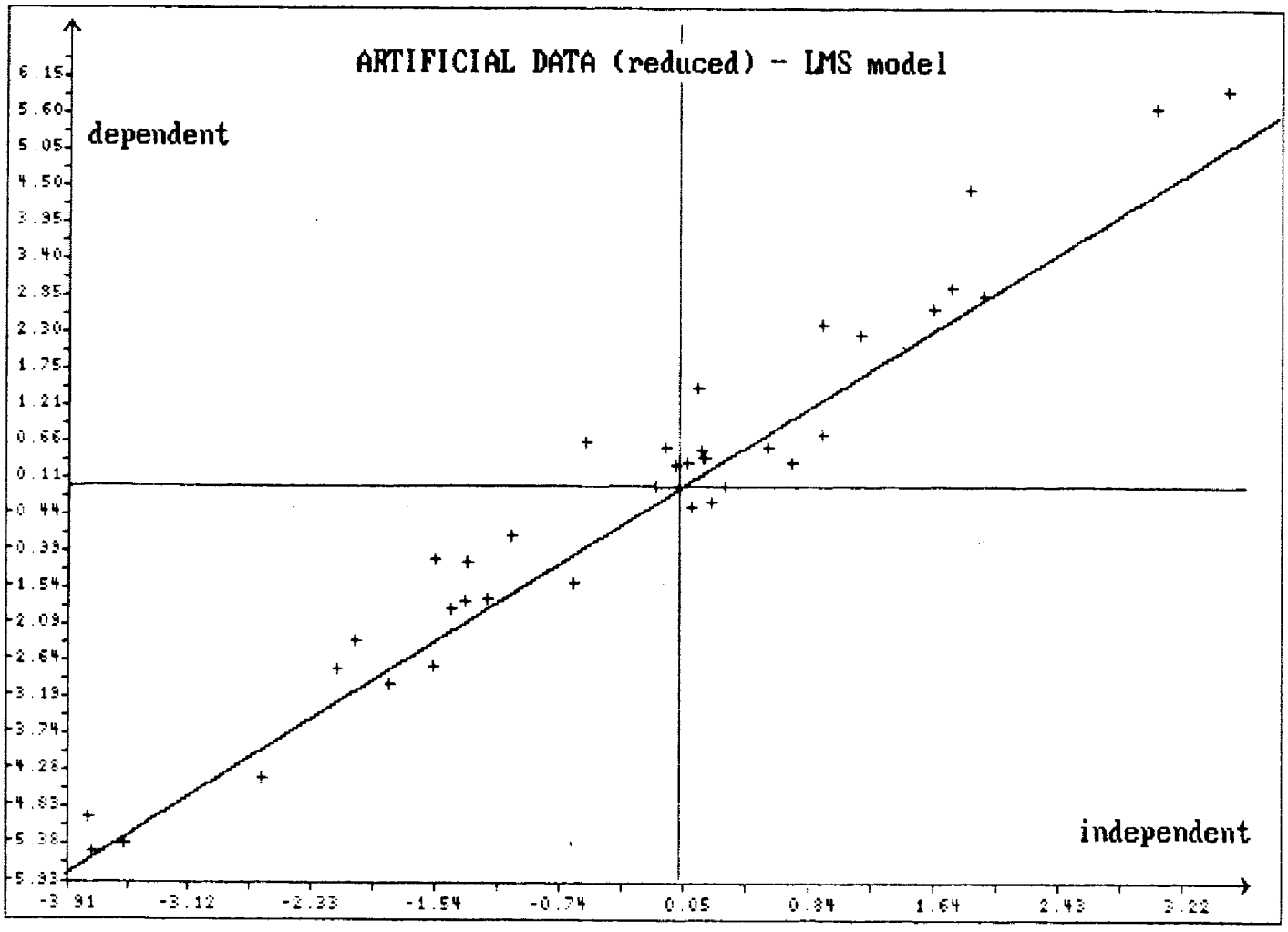
ARTIFICIAL DATA (y=x+e; x..U(-4,4); e..N(0,1.3))

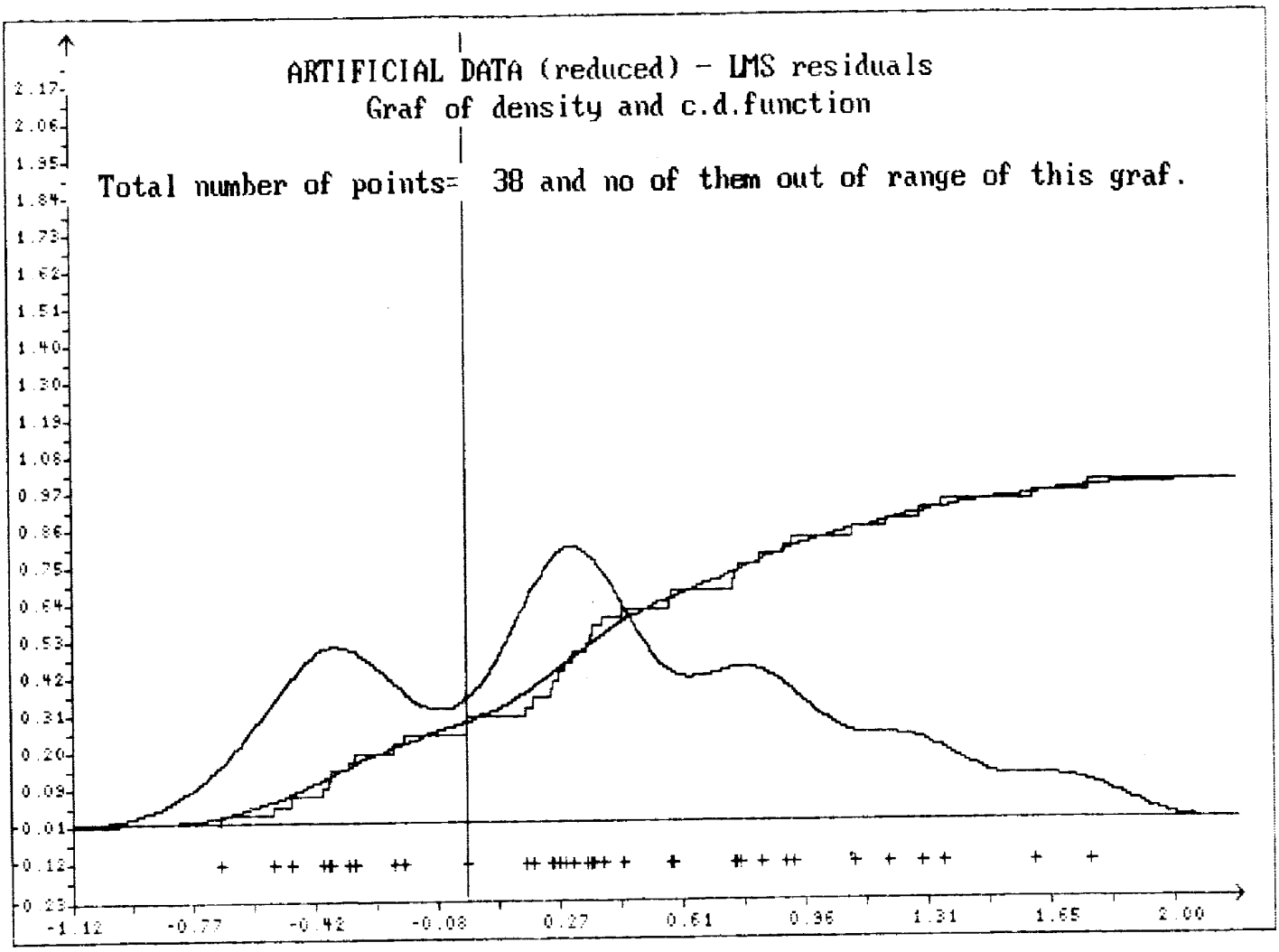
****

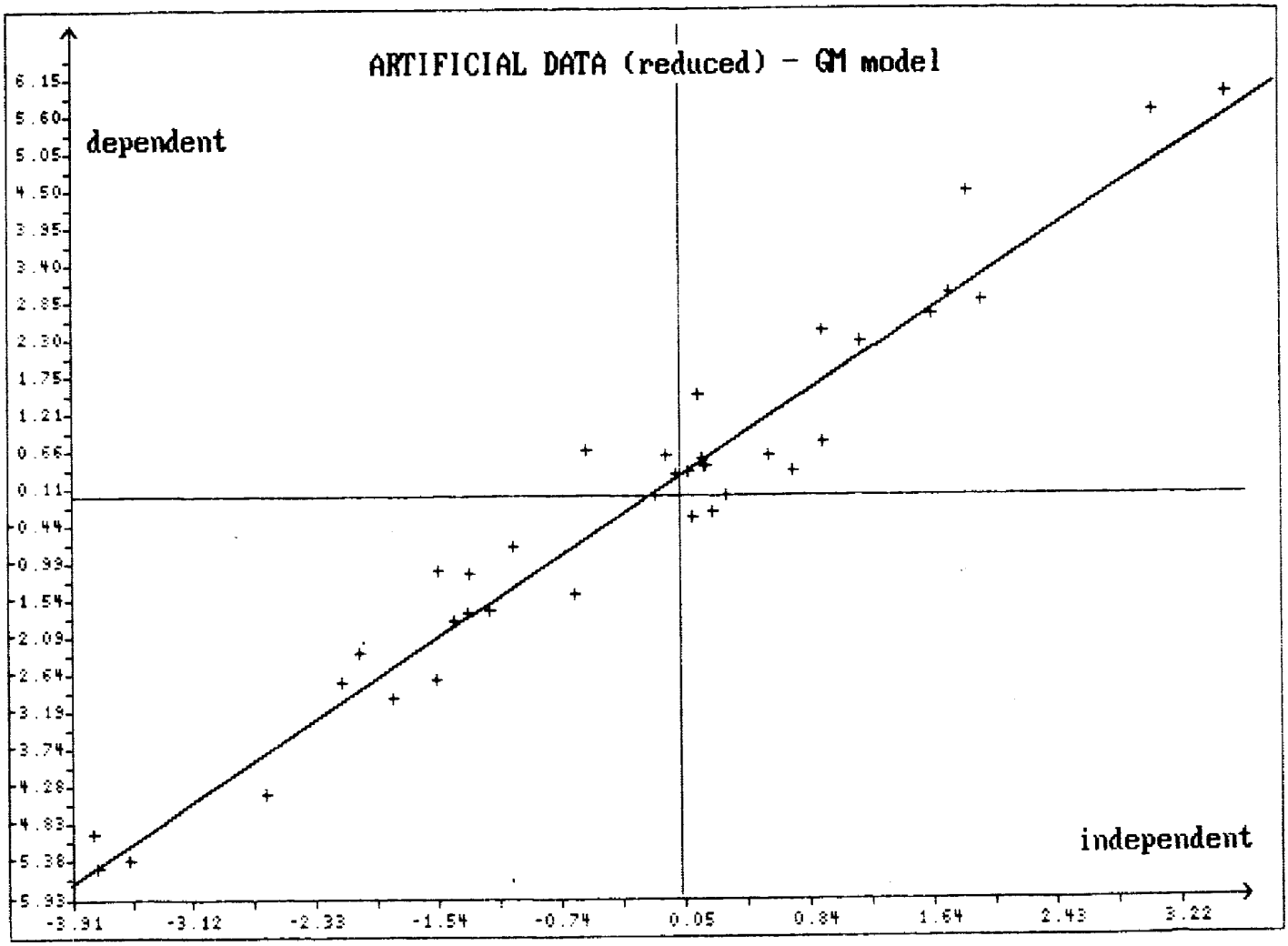
case	x	y	case	x	y
1	-3.758713	-5.450430	20	3.018787	5.729336
2	-1.443068	-1.807758	21	1.925544	2.900774
3	-1.069856	-0.705167	22	0.207729	-0.176134
4	-1.215064	-1.653717	23	-0.017527	0.366336
5	0.147676	0.586328	24	1.153117	2.315358
6	-2.664302	-4.326543	25	1.724177	3.034948
7	-3.780872	-4.921943	26	-0.592642	0.711471
8	-0.665691	-1.383678	27	3.484233	6.008646
9	-0.089969	0.628017	28	0.912645	0.830367
10	-1.564029	-2.658770	29	0.562032	0.644310
11	-1.840087	-2.936874	30	0.117101	1.522670
12	-2.173251	-2.683591	31	1.614859	2.711646
13	-1.549990	-1.039650	32	1.842544	4.528104
14	-1.357830	-1.690449	33	0.303299	0.055823
15	0.159062	0.491681	34	-1.344843	-1.097367
16	-3.551338	-5.326357	35	0.717807	0.389267
17	-2.058207	-2.255504	36	0.167538	0.496077
18	-0.144130	0.053167	37	0.080927	-0.255941
19	0.912632	2.469360	38	0.055200	0.392068

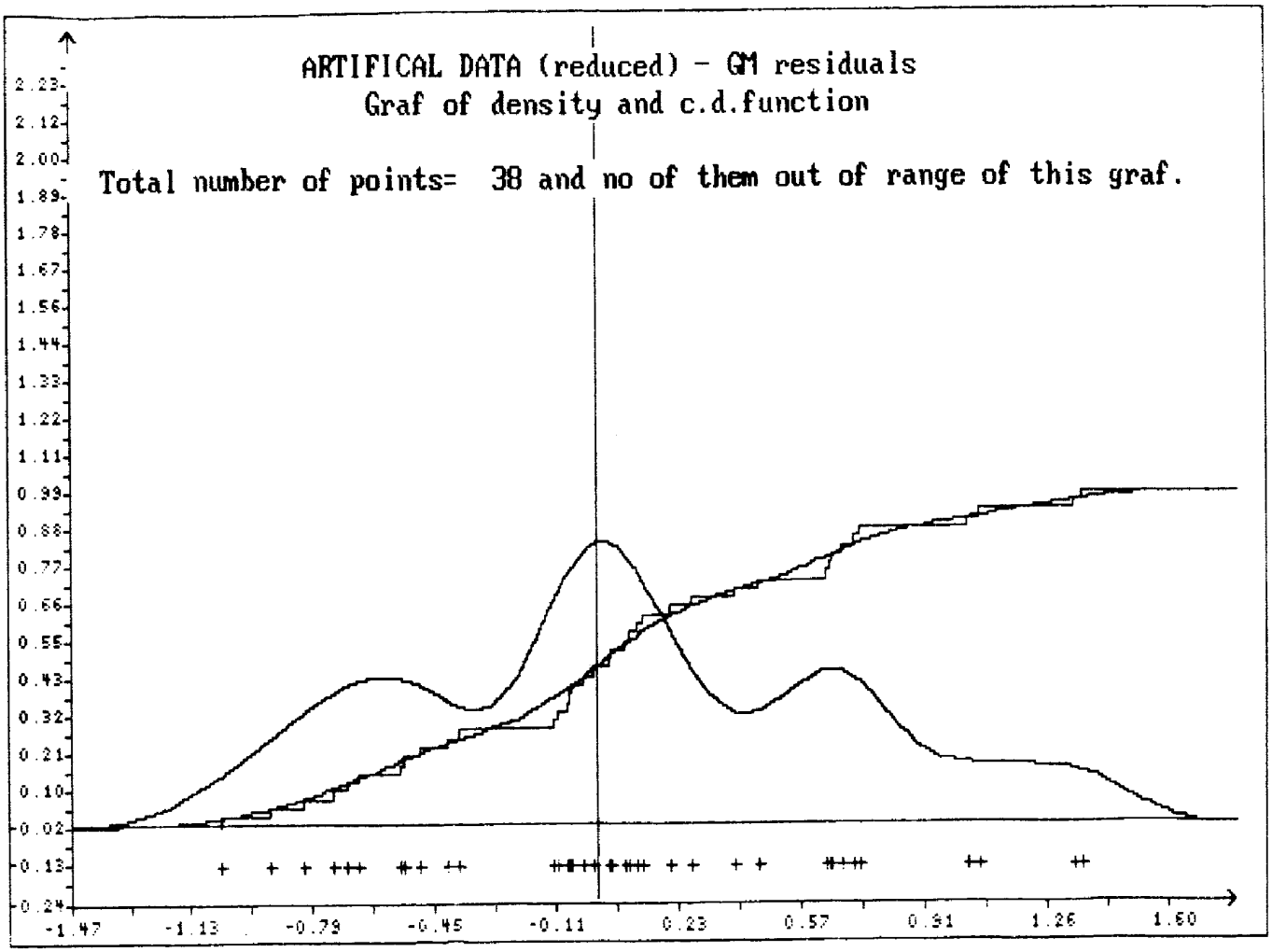


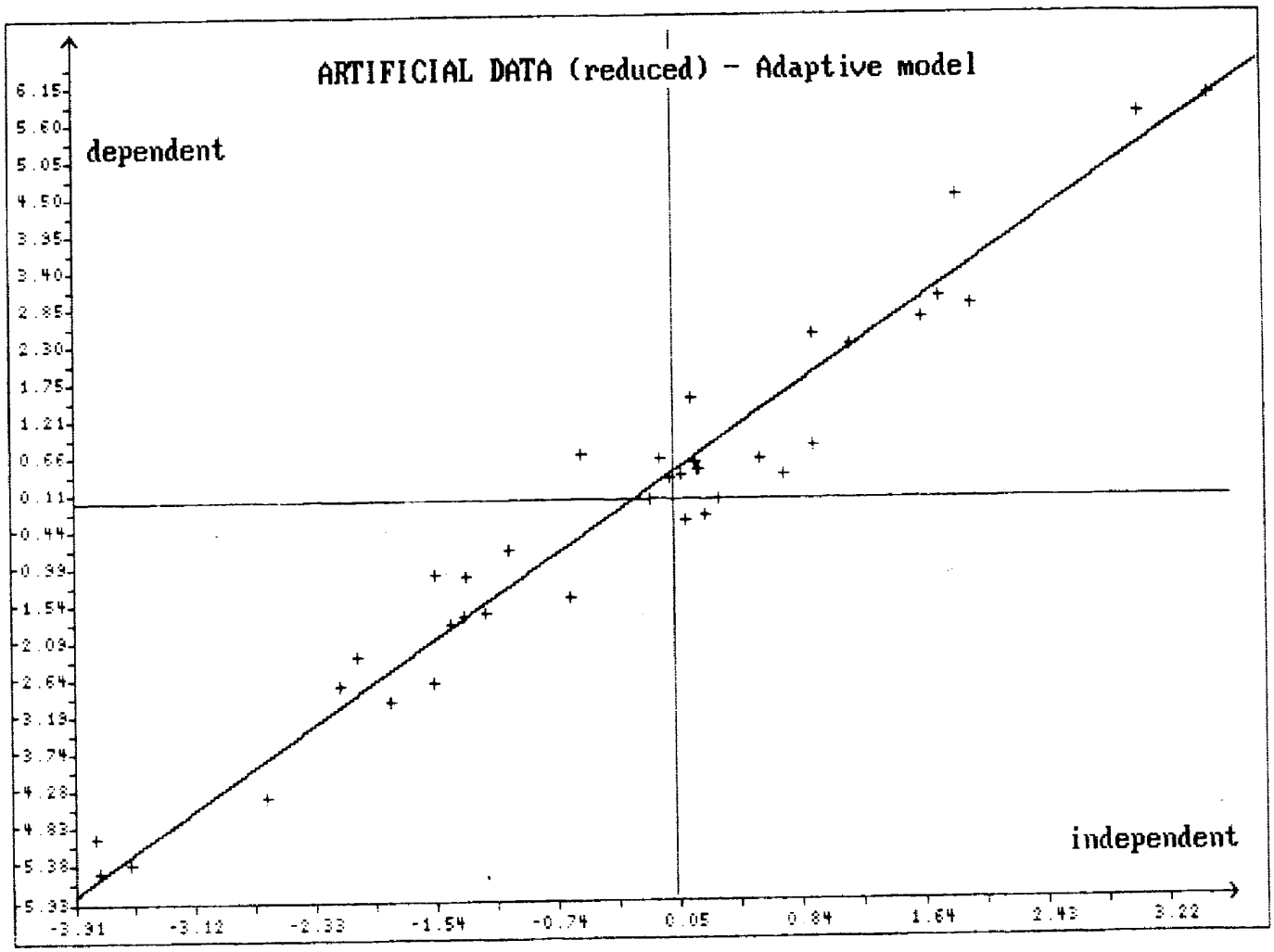

ARTIFICIAL DATA (reduced) - LS residuals
Graf of density and c.d.function
Total number of points= 38 and no of them out of range of this graf.

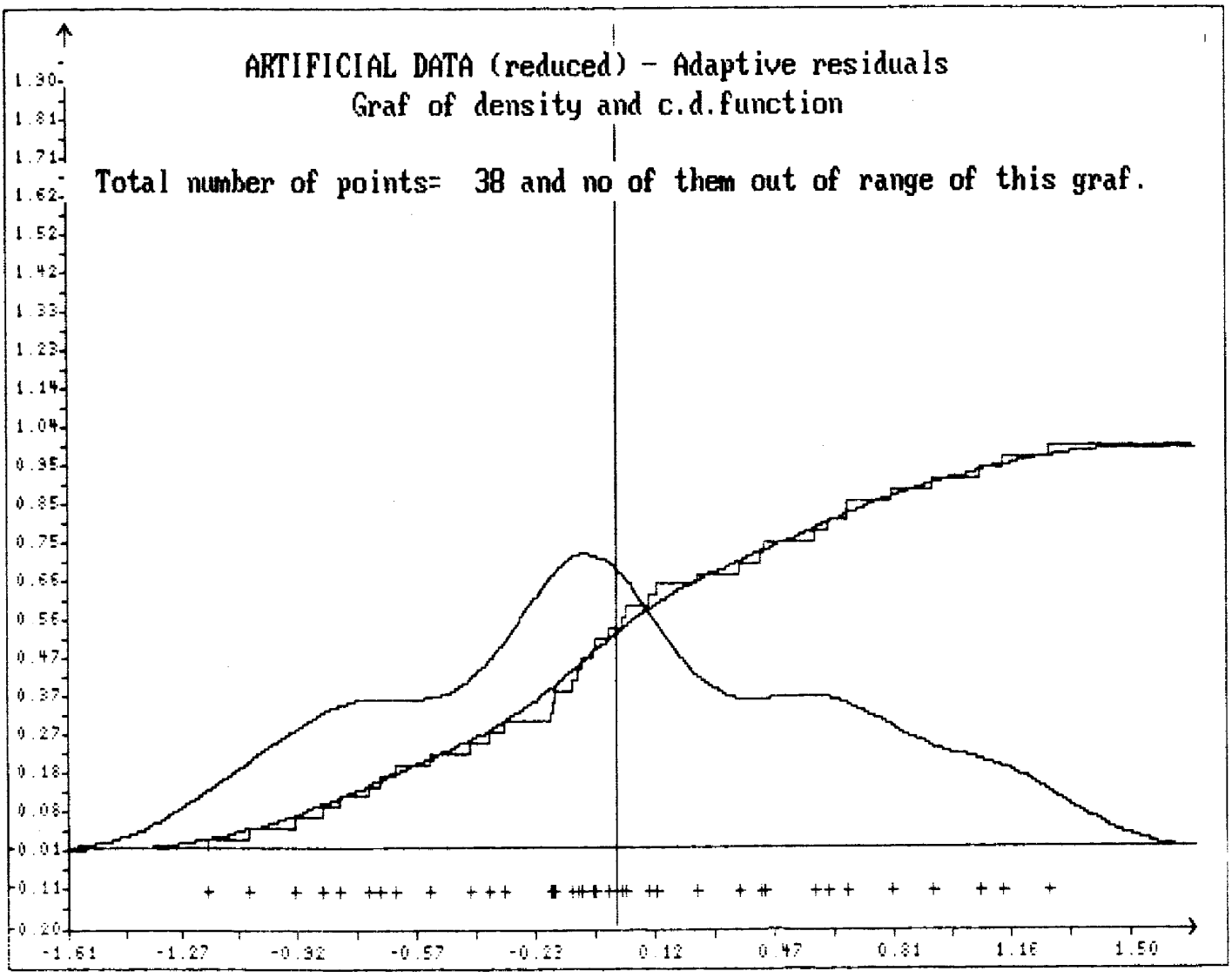

ARTIFICIAL DATA (reduced) - LMS model
dependent
independent


ARTIFICIAL DATA (reduced) - LMS residuals
Graf of density and c.d.function
Total number of points= 38 and no of them out of range of this graf.
-1.12
-0.77
-0.42
-0.08
0.27
0.61
0.96
1.31
1.65
2.00


ARTIFICIAL DATA (reduced) - GM model
dependent
independent
-3.91
-3.12
-2.33
-1.54
-0.74
0.05
0.84
1.64
2.43
3.22


ARTIFICAL DATA (reduced) - GM residuals
Graf of density and c.d.function
Total number of points= 38 and no of them out of range of this graf.
2.23
2.12
2.00
1.89
1.78
1.67
1.56
1.44
1.33
1.22
1.11
0.99
0.88
0.77
0.66
0.55
0.43
0.32
0.21
0.10
-0.02
-0.13
-0.24
-1.47
-1.13
-0.79
-0.45
-0.11
0.23
0.57
0.91
1.26
1.60

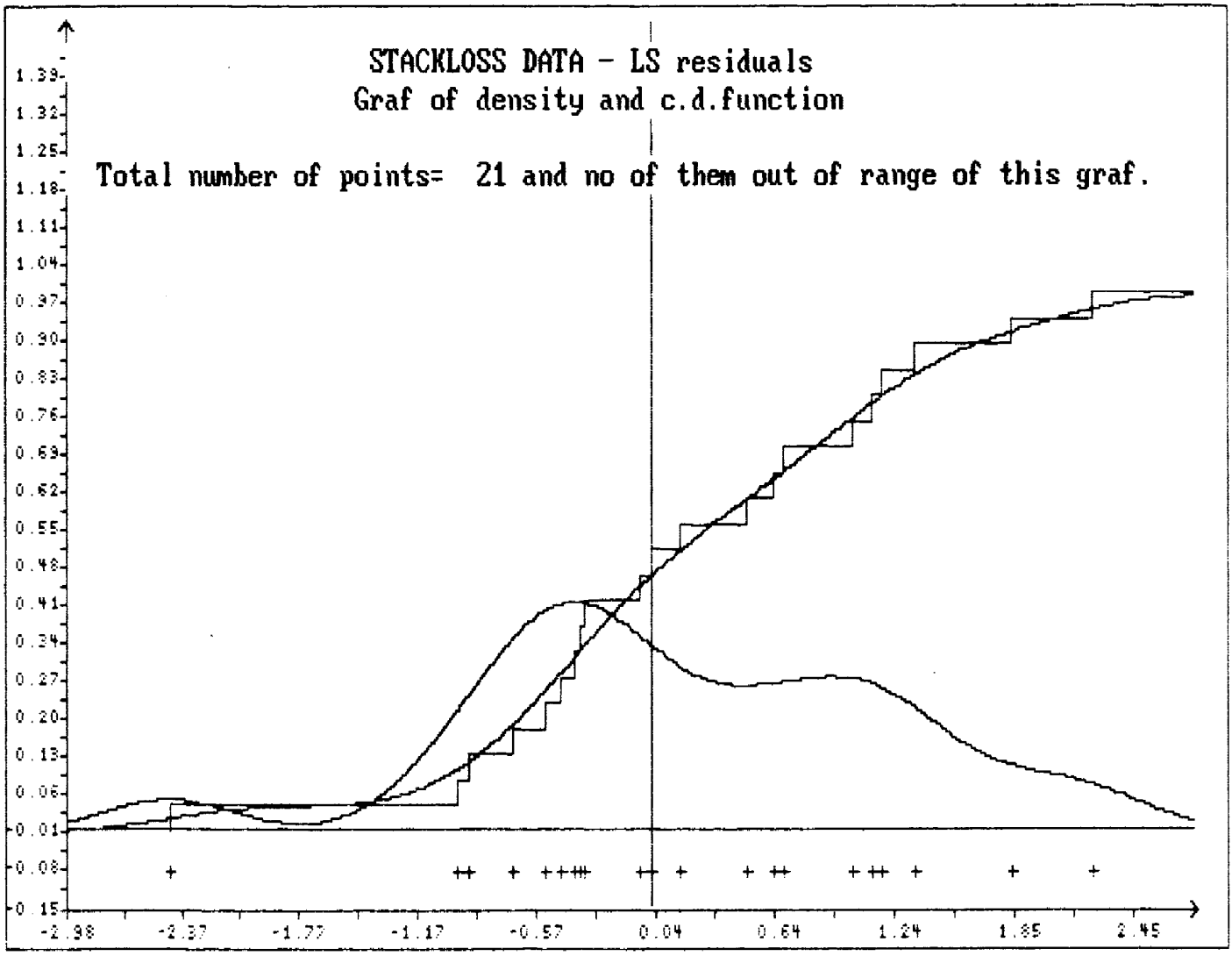

ARTIFICIAL DATA (reduced) - Adaptive model
dependent
independent
6.15
5.60
5.05
4.50
3.95
3.40
2.85
2.30
1.75
1.21
0.66
0.11
-0.44
-0.99
-1.54
-2.09
-2.64
-3.18
-3.74
-4.28
-4.83
-5.38
-5.93
-3.91
-3.12
-2.33
-1.54
-0.74
0.05
0.84
1.64
2.43
3.22

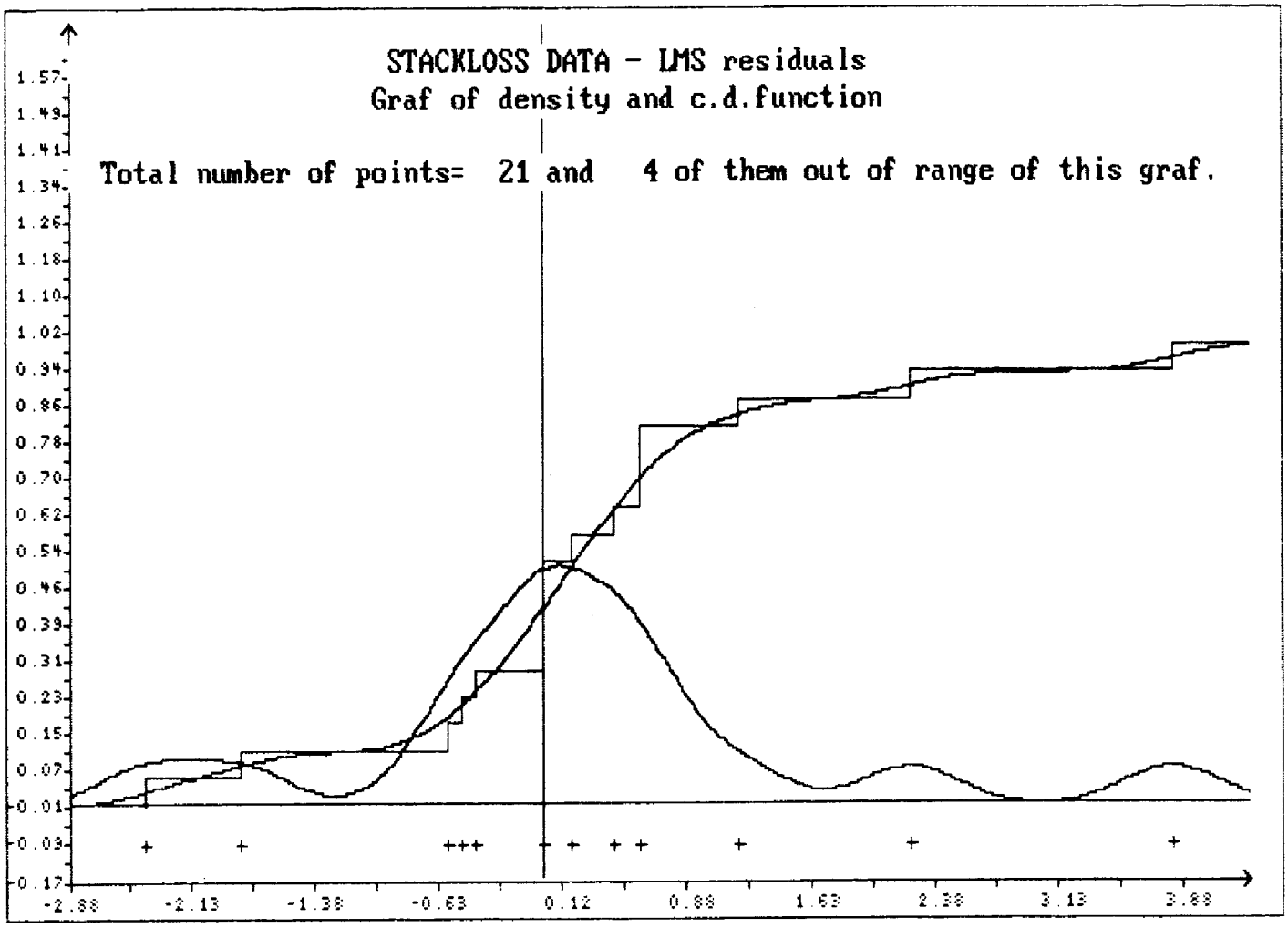


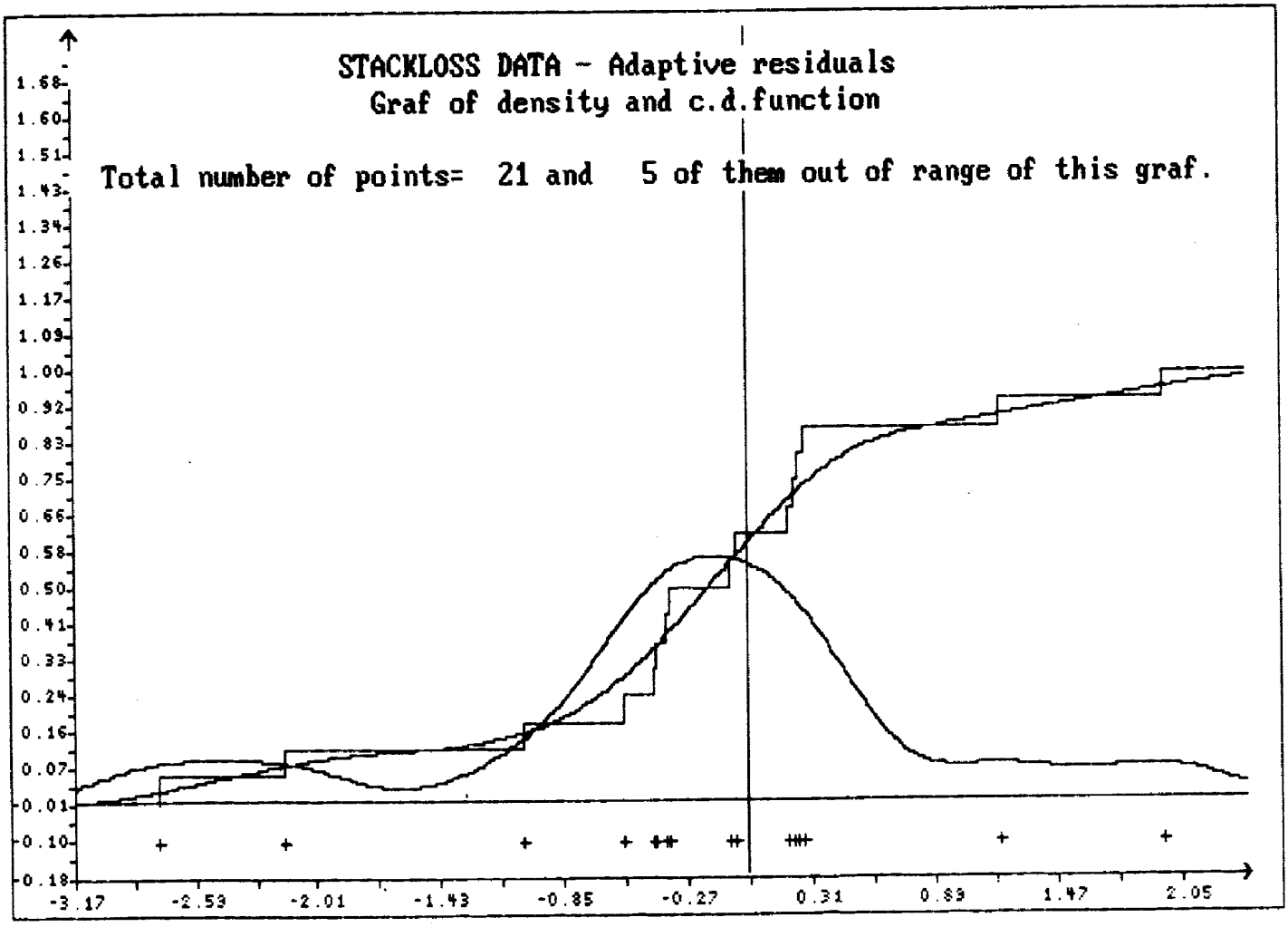
The second example will consider Stackloss data (see Brownlee (1965)). Let us explain abbreviation in the next table. LS denotes again Least Square estimate, $\hat{\beta}(.5)$-regression quantiles for $\alpha = .5$, $\hat{\beta}_{PE}(.10)$ - Trimmed Least Squares where trimming was symmetric and trimmed off the 10 % of largest and smallest residua in the model with preliminary estimator which was in this case average of the αth and $(1-\alpha)$th regression quantiles (i.e. $\hat{\beta}_{preliminary} = (\hat{\beta}(\alpha) + \hat{\beta}(1-\alpha))/2$ for $\alpha = .1$), $\hat{\beta}_{KB}(.15)$ - Least Square estimate after trimming off points according to regression quantiles $\hat{\beta}(.15)$ and $\hat{\beta}(.85)$, Huber and Andrews -, corresponding M-estimates, LMS - Least Median of Squares (in fact model in which $[(n+p-1)/2]$th order statistic of residua was minimized), LTS (Rousseeuw) - Least Trimmed Squares (in fact this estimate is $\hat{\beta}_{PE}(\alpha)$ where as the preliminary estimator served LMS) and Adaptive - adaptive estimator of Hellinger type.

STACKLOSS DATA

Method	*Estimates of coefficients*			
	Intercept	*Air Flow*	*Temperature*	*Acid*
LS	39.92	-.72	-1.30	.15
$\hat{\beta}(.5)$	39.69	-.83	-.57	.06
$\hat{\beta}_{PE}(.10)$	40.37	-.72	-.96	.07
$\hat{\beta}_{KB}(.15)$	42.83	-.93	-.63	.10
Huber	41.00	-.83	-.91	.13
Andrews	37.20	-.82	-.52	.07
LMS	34.50	-.71	-.36	.00
LTS (Rouseeuw)	35.48	-.68	-.56	.01
Adaptive	34.50	-.72	-.36	.00




STACKLOSS DATA - LMS residuals
Graf of density and c.d.function
Total number of points= 21 and 4 of them out of range of this graf.
1.57
1.49
1.41
1.34
1.26
1.18
1.10
1.02
0.94
0.86
0.78
0.70
0.62
0.54
0.46
0.39
0.31
0.23
0.15
0.07
0.01
-0.09
-0.17
-2.88
-2.13
-1.38
-0.63
0.12
0.88
1.63
2.38
3.13
3.88


STACKLOSS DATA - Adaptive residuals
Graf of density and c.d.function
Total number of points= 21 and 5 of them out of range of this graf.
1.68
1.60
1.51
1.43
1.34
1.26
1.17
1.09
1.00
0.92
0.83
0.75
0.66
0.58
0.50
0.41
0.33
0.24
0.16
0.07
-0.01
-0.10
-0.18
-3.17
-2.59
-2.01
-1.43
-0.85
-0.27
0.31
0.89
1.47
2.05

References

Beran, R. (1978): An efficient and robust adaptive estimator of location, AS 6 pp. 292-313.

Bickel, P. J. (1975): One-step Huber estimates in the linear models, J. American Statist. Ass. 70 pp. 428-433.

Bickel, P. J. (1982): The 1980 Wald Memorial Lecture: On adaptive estimation, Ann. Statist. 10, pp. 647-671.

Brownlee, K. A. (1965): Statistical Theory and Methodology in Science and Engineering, 2nd ed., John Wiley & Sons, New York.

Fisher, R. A. (1922): On the mathematical foundations of theoretical statistics, Philos. Trans. Roy. Soc. London Ser. A 222, 309-368.

Hampel, F. R.; Ronchetti, E. M.; Rousseeuw, P. J.; Stahel, W. A. (1986): Robust Statistics, The Approach Based on Influence Functions, J. Wiley & Sons, New York.

Huber, P. J. (1969): Théorie de L'inference Statistique Robuste, Les presses de l'Université de Montreal.

Jeffreys, H. (1939): Theory of Probability. Clarendon Press, Oxford.

Jurečková, J. (1983): Robust estimation of location and regression parameters and their second order asymptotic relations, Proc. 9th Prague Conf. on Inform. Theory, pp. 19-32, Reidel.

Jurečková, J. (1977): Asymptotic relations of M-estimates and R-estimates in linear regression model, Ann. Statist. 5, pp. 464-472.

Jurečková, J. and Sen, P. K. (1981): Sequential procedures based on M-estimators with discontinuous score function, J. Statist. Planning and Inference 5, pp. 253-266.

Jurečková, J. (1984): Regression quantiles and trimmed least squares estimators under a general design, Kybernetika 20, pp. 345-357.

Koenker, R. and Bassett, G. (1978): Regression quantiles, Econometrika 46, pp. 33-50.

Novovičová, J. (1990): M-estimators and gnostical estimators for identification of regression model, Automatica 26, May, 1990 (to appear).

Pearson, K. (1902): On the mathematical theory of errors of judgement, with special reference to the personal equation, Philos. Trans. Roy. Soc. Ser. A 198, 235-299.

Rousseeuw, P. and Yohai V. (1984): Robust regression by means of S-estimates, in Robust and Nonlinear Time Series Analysis (eds. J. Franke, W. Härdle and D. Martin), Lecture Notes in Statistics No. 26, Springer Verlag, pp. 256-272.

Ruppert, D. and Carroll, R. J. (1980): Trimmed least squares estimation in the linear model, J. American Statist. Ass. 75, pp. 828-838.

Stein, C. (1956): Efficient nonparametric testing and estimation, Proc. Third Berkeley Symp. Math. Statist. Prob. University of California Press, 1 pp. 187-196.

Student (1927): Error of routine analysis, Biometrika 19, 151-164.

Stone, C. (1975): Adaptive maximum likelihood estimation of location parameter. Ann. Statist. 3 pp. 267-284.

Víšek, J. Á. (1983): Sensitivity of the test risk with respect to contamination, Commun. Statist. - Sequential Analysis 2/3, pp. 243-258.

Víšek, J. Á. (1986): Sensitivity of the test error probabilities with respect to level of contamination in general model of contaminacy, J. Statist. Planning and Inference 14, pp. 281-299.

Víšek, J. Á. (1989): Estimation of contamination level in the model of contaminacy with general neighbourhoods, Kybernetika 25, pp. 278-297.

Víšek, J. Á. (1990a): Adaptive estimation in linear regression model, research report ÚTIA ČSAV

Víšek, J. Á. (1990b): Adaptive maximum-likelihood-like estimation of regression model, research report ÚTIA ČSAV.

Yohai, V. J. (1974): Robust estimation in the linear model, Ann. Statist. 2, pp. 562-567.