Diagnosis of some model deficiencies using recursive residuals

Douglas M Hawkins
University of Minnesota
Technical Report No. 501
September, 1987

Diagnosis of some model deficiencies using recursive residuals

Douglas M Hawkins
Department of Applied Statistics
University of Minnesota
ST PAUL
MN

Abstract

Ordinary least squares residuals are the most common diagnostics for departures from the model assumptions of regression, but may completely miss some departures from model, and provide equivocal diagnoses of others. Recursive residuals are much more effective under circumstances where the model holds at one of the ends of the data set, and are even more effective used in conjunction with the a recently-developed technology - that of 'self-starting' cusums. Diagnosis using recursive residuals and self-starting cusums, and that using OLS residuals and the more classical forms of the cusum are compared in four data sets involving departures from model, and it is shown how in each the new methods provide a clearer more reliable diagnosis of the departure.

Keywords

Cusum techniques, Outliers, Change of regime, Heteroscedasticity

Introduction - Ordinary and Recursive Residuals

Consider the linear regression model

$$Y = X\beta + \epsilon,$$

(X of order n×p) whose conventional distributional assumptions are that the true errors ϵ_i are independently and identically distributed normal variates.

The most common tests for the fit of the model are based on the fitted ordinary least squares (OLS) residuals

$$e = Y - X\hat{\beta}$$

which, as is well known, can also be written

$$e = (I - H)\epsilon$$

where H is the 'hat' projector matrix $X(X^TX)^{-1}X^T$.

Under the model, $e \sim N[0,\sigma^2(I-H)]$

The OLS residuals have a number of deficiencies as diagnostics of departure from model. Among the general deficiencies are that they are

1 Not of equal variance, even where the true residuals are,
2 Mutually correlated,
3 Closer to normally distributed than the true residuals, as a result of the operation of the central limit theorem.

Recursive residuals are defined by an iterative process. Start with the full sample, and compute the OLS residuals. The last of these, e_n, is distributed under the model as $N[0,\sigma^2(1-h_{nn})]$. From it, define a scaled quantity

$$r_n = e_n/\sqrt{(1-h_{nn})} \sim N(0,\sigma^2).$$

Having computed r_n, remove the nth observation from the sample, and

repeat the entire process, duly computing another $N(0,\sigma^2)$ quantity r_{n-1} say. Continue stripping the sample away observation by observation, at each stage computing a fresh scaled residual r_i for the last of the remaining observations in the data set. As only n-p of the observations can be deleted in this way without singularity setting in, only n-p such scaled residuals can be defined. They are called the recursive residuals. Let the recursive residuals be written as a column vector r say.

Under the model, $r \sim N(0,\sigma^2 I)$, and so the recursive residuals avoid the first two of the general deficiencies of OLS residuals; as we shall see, while they share the third deficiency they suffer from it less.

The recursive residuals have the property $\Sigma_i r_i^2 = \Sigma_i e_i^2$ - ie they reproduce the residual sum of squares of the regression. They do not however have any property analogous to that of OLS residuals that $\Sigma_i e_i x_{ij} = 0$ for every predictor j.

Recursive residuals have a long history. Their properties as diagnostics for departures from model were set out by Brown, Durbin and Evans (1975), as well as the discussants of that paper. Despite this and the attractions mentioned above (though perhaps partly because a casual reader of Brown et al could leave with the false impression that they are applicable only to time series data), recursive residuals do not seem to be widely used for diagnosis of departures from the model specification in multiple regression - for example, they are not among the standard diagnostics produced by most package programs. The purpose of this note is to give a few comparative examples of the performance of diagnostics based on recursive and on OLS residuals to illustrate the strengths and weaknesses of each.

The computation of recursive residuals requires some care and attention

to the numerical analysis aspects. As the successive points are deleted, the conditioning of the design matrix deteriorates steadily, and this fact makes the use of numerically stable methods essential. An attractive approach is that based on triangular factorization, for which numerically stable methods of updating and downdating are given, for example, in Gragg, Leveque and Trangenstein (1979). The computation may be done by either downdating (starting with the full data set and removing cases sequentially) or updating (starting with the final set of size p and successively adding the rest of the sample case by case). Of the two, the updating approach is the more stable and so is computationally preferable. Code in BASIC for stable updating and downdating may be found in Maindonald (1984) where the term 'Givens residual' is used for the recursive residual.

In addition to this issue of numeric precision, it is a mathematical requirment for computation of the recursive residuals that the intermediate design matrices all be of full rank. Quite often, in the deletion process one comes to a point where a particular case is required to preserve rank. When this occurs, that case is kept in the current basis, and the next case deleted in its stead. Provided there is at least one nonsingular square submatrix of X, the usual result is a set of n-p recursive residuals (usually but not necessarily corresponding to the last or 'rightmost' n-p cases) and a set of p cases which are retained in all regressions (usually but not necessarily the first or 'leftmost' p cases).

Apart from the computational issue of whether they are computed by updating or downdating, there is also the issue that recursive residuals can be computed either 'backward' or 'forward' - that is either by deleting observations successively from right to left as described above ('backward'), or from left to right ('forward'). While either direction will produce the same sum of squares for the recursive residuals, the actual values of the 'backward' and the 'forward' recursive residual at

any point for which both are defined are from regressions based on non-intersecting subsets of the data and may be completely different, so that the correct diagnosis of the deficiency may require the computation and analysis of both sets of recursive residuals.

Effect of departures from model

Because of the way the true residuals are smeared across the fitted residuals by the projector I-H (which can have substantial off-diagonal elements), the ordinary residuals suffer from further deficiencies not shared by recursive residuals when used to diagnose certain types of departure from model specification. Examples of this type are situations in which the usual regression assumptions apply in a portion of the data set, but not in the whole data set. Under these circumstances, all of the ordinary residuals are contaminated by the lack of fit of a portion of the true residuals, and so may fail to present a clear picture of the model departure. The recursive residuals however avoid this difficulty provided the departure from model does not continue to the 'left end' of the data set, since once the contaminated portion of the data has been deleted, the remaining recursive residuals are 'clean'. Comparison of the recursive residuals computed later in the computation with those computed earlier then shows the problem up clearly.

Amplifying this point of the general behaviour of the two types of residual under departures from model, let us look briefly at the effect of some common departures from model on OLS and recursive residuals. Notable are:-

1 Outliers. There is a vast literature on outliers in regression which we will not cover again here. The projection going from ϵ to e has the well-known effect of (i) attentuating the real departure

of an outlier from the regression; and also (ii) inducing apparent deviation in inliers. In the extreme case, one can find the OLS residuals providing completely wrong diagnoses of which points are inlying and which outlying as is illustrated below. This defect is clearly avoided by the recursive residuals provided at some stage there comes a point at which all outliers have been excluded and the remaining cases conform to the model.

Interestingly though, a correct diagnosis may result even if some outliers are never entirely excluded in the computation of the recursive residuals. This can happen because the retained outliers bias the regression function, which may give the fitted recursive residuals a nonzero overall mean. Thus a check of the mean of the recursive residuals (unlike that of the OLS residuals which is necessarily zero) can be very informative.

2 Nonlinearity. Suppose for illustrative purposes that the true regression of Y on one of the x is a convex quadratic rather than linear, and suppose also that the data set happens to be ordered on that x. The plot of the OLS residuals on x will then tend to be U shaped, something that is not always easy to see in a residual plot. The recursive residuals behave quite differently. As each point is deleted, the resulting fitted line slopes decreases, with the result that all the recursive residuals tend to have positive expectations. Both visually on an index plot and more formally from the mean of the recursive residuals, this tendency is much easier to detect.

We have now mentioned two departures from model that can give an overall nonzero mean recursive residual, but this second departure is distinguished from the first by the fact that both the 'backward' and the 'forward' recursive residuals have a non-zero mean of the same sign, whereas in the case of never-deleted outliers one will tend to see a clear outlier syndrome for one direction of

computation, and the non-zero mean recursive residual for the other. Thus a comparison of the forward and backward recursive residuals will show whether there is a problem of non-linearity or one of outliers at the end of the sequence.

3 Heteroscedasticity. It is clear that the OLS residuals are a poor indicator of heteroscedasticity when the larger variance occurs at points of higher leverage - this is because of the $(1-h_{ii})$ factor in the variance of e_i. One by-product of the standardisation involved in defining the recursive residuals is the removal of this effect, so that the recursive residuals are not blind to higher variance on high leverage points.

4 Omitted predictor. In general, an omitted predictor is undetectable by any diagnostics unless it happens to relate to something in the data set. Consider the values of the omitted predictor adjusted for the predictors in the regression. If these adjusted values vary reasonably smoothly with the order of the cases, then the omission may be seen in the form of a slowly varying non-zero mean in the residuals (both OLS and recursive), and this may be detected with suitable checks on the residual's current mean such as the cusums discussed below. Another possibility is that the adjusted values may be more variable in some portions of the ordered sequence than in others; in this case the omission will show up in the form of non-constant smoothly changing variance. This may be detected by diagnostics for general changes in the variance from case to case. The implication of this is that a slowly drifting mean or variance of the OLS or recursive residuals could be an indicator of an omitted predictor.

5 Change of regime. If any regression coefficient changes from one portion of the data to another, then the recursive residuals will show a clear 'two-phase' behaviour. Those from the end will consist

of recursive residuals from a single phase, and so will be $N(0,\sigma^2)$ while those from the earlier stages of deletion will show nonzero mean, increased variance or both. Again diagnostics for location and scale of the recursive residuals will show up the deficiency.

6 Non-normal errors. As already mentioned, the OLS residuals suffer from a 'central limit' effect in that the fitted residuals are closer to normal than the true residuals. This effect is also present in the recursive residuals, but is somewhat attenuated by the fact that as one computes the successive recursive residuals, each involves fewer of the original ϵ_i.

Ordinary and self-starting cusums

As these comments illustrate, the diagnosis of many of the common departures from model may be made by detecting systematic but otherwise quite general departures from zero mean and constant variance in the ordinary and/or the recursive residuals. Powerful diagnostics for each may be made using cumulative sum techniques (cusums). Consider a series of data say $Z_i \sim N(\mu,\sigma^2)$, where μ and σ are known. Conventional cusums for normal data (either explicitly or implicitly) standardise these values to $U_i=(Z_i-\mu)/\sigma$ and then apply the known distribution theory of partial sums of standard normal quantities to set up monitoring schemes of known average run length.

The standard cusums are for location, and for scale. A location cusum normally consists of the running total (or something mathematically equivalent) of the U_i.

There have been several proposals for cusums to check for constancy of variance. The older and better known are those based on the cusum of squares of the putatively $N(0,1)$ quantities U_i, an approach which has

the disadvantages of requiring development of its own inferential theory, of high sensitivity to outliers, and of not being very easily related to the cusum of the U_i.

An alternative proposal (the 'square root scale cusum') is that of Hawkins (1981), and involves computing the quantity

$$V_i = (\sqrt{|U_i|}-0.822)/0.345$$

whose distribution is very close to N(0,1) if the U_i are N(0,1), but whose mean changes upward or downward if the variance of the U_i increases or decreases. A cusum of the V_i therefore provides a mechanism by which scale changes in the original data can be monitored using an ordinary location cusum of N(0,1) quantities. This procedure provides (at a cost of some loss of power) a method of controlling for scale that is more robust than that of running sums of squares, and that uses the well-known procedures for location cusums of normal quantities. As an extra benefit of presentation, one may plot the cusums of U and V together on the same graph (since they have identical distributions and therefore the same action limits).

This combination of cusums of the U_i to check for mean and V_i to check for variance provides a very effective simultaneous check on the constancy of the data.

Problems with this traditional approach to cusums arise when the parameters μ and/or σ are not known. Standard practice has been to substitute for these parameters sample estimates, but this creates difficulties. Not only is the distribution of the resulting standardised quantities U_i Student's t rather than normal, but a correlation is also induced between them. These two departures from the asymptotic assumptions obviously affect the run length distribution of the cusum, but little work seems to have been done quantifying the effect.

A different approach addressing these problems is the 'self-starting' cusums developed in Hawkins (1987). These cusums are designed for the situation in which μ and σ are not known, and learn about the unknown parameters as they go, simultaneously providing cusum control on the constancy of the process. The procedure set out in Hawkins (1987) was for the case neither μ nor σ known, while under the regression model, the recursive residuals r_i are independent $N(0,\sigma^2)$ with σ^2 the only unknown parameter, the mean being zero under the model. To deal with this additional information, we modify the approach slightly - define:

The running standard deviation $s_i = \sqrt{\left[\sum_{j=1}^{i} r_j^2/i\right]}$ $i = 1, 2....$

$T_i = r_i/s_{i-1} \sim t_{i-1}$, $i = 2, 3...$

$$U_i = \frac{8i-7}{8i-5} \log\left[1 + \frac{T_i^2}{i-1}\right]$$

The successive T_i are t distributed with i-1 degrees of freedom and mutually independent. Each tests the corresponding recursive residual r_i for zero mean, studentising using the running standard deviation of all previous recursive residuals. The transformation of T_i to U_i provides a quantity that is (Peizer and Pratt 1968) very close to N(0,1).

By cusumming the U_i and the derived scale check quantities V_i, one can then check the successive recursive residuals r_i for departure from zero mean and constant (but unknown) variance. We term these cusums 'self-starting' since, unlike the conventional ones, they do not require any starting information in the form of known parameter values. Unlike the procedure obtained by plugging sample estimates into the conventional cusum for known parameters, the approach using self-starting cusums is distributionally correct apart from the quite minor approximation involved in the normalising transformation from the t distribution to N(0,1).

Another approach using some of these distributional properties is that

of Hester and Quesenberry (1984). These authors also make use of the independence and t distribution of the T_i, but rather than normalising prefer to transform them to uniform by the conditional probability integral transform of O'Reilly and Quesenberry (1973) and from there to scores distributed as x^2_2. These they then use to form an omnibus test for monotonic heteroscedasticity. Their approach is designed for the specific situation of a monotonically changing variance though, and seems less likely to perform well against departures from different aspects of the model and or where the departure is not monotonic across cases.

It may be noted that T_i is the optimal single-outlier regression test statistic for testing case i in the set $\{1,2,\ldots,i\}$. The individual values of the T_i are thus of particular interest as being good single-outlier diagnostics, particularly if any outliers happen to be located in the portion of the sequence deleted relatively early.

Diagnostics used for checking the recursive residuals

In principle, any of the diagnostic checks that are carried out using OLS residuals could be applied also using recursive residuals. Two that are particularly appropriate are (i) a normal probability plot, and (ii) location and scale cusums. The former provides a good indication of a number of departures from model - not only of the normality part of the assumption, but also changes of regime, outliers, omitted variables and heteroscedasticity.

There have been a number of previous papers on the use of cusums for detection of departures from model. Apart from the landmark paper by Brown et al proposing the use of cusums of the recursive residuals and their squares, there is the critical evaluation by Garbade (1977) of the cusum of squares of recursive residuals. McCabe and Harrison suggest

the use of the cusum of squares of the OLS residuals, while Galpin and Hawkins (1984) discuss the use of the location and 'square root scale' cusums of the recursive residuals. In all of these papers however, the problem of unknown parameters is dealt with by 'plugging in' an estimate of σ. Not only does this involve the difficulties of violating the assumptions of exact N(0,1) distribution (as mentioned above) it also creates difficulties of interpretation in the event of a departure from model that, since the residual standard deviation is itself contaminated by the departure from model, the resulting cusum diagnostics may be misleading about the true nature of the departure.

Self-starting cusums provide a much more attractive analysis for recursive residuals. If the recursive residuals were constructed by deleting points from right to left, the cusums are constructed by accumulating from left to right. This has the implication that if there is a departure from model affecting only a portion of the data, then the self-starting cusums automatically learn about σ from the uncontaminated portion of the data, and thereby perform effectively in diagnosing the departure from model in the later part of the series correctly.

The direction of computation of a self-starting cusum matters, and must match the direction of computation of the recursive residuals - 'backward' recursive residuals must be cusummed forward, and 'forward' recursive residuals must be cusummed backward. If one takes a series of data and runs it through a conventional cusum backwards, the plot obtained is a mirror image of that obtained by computing it forwards. But because of their use of a running standard deviation for studentisation, this is not true of self-starting cusums, where the backward run of a set of data may produce quite different plots from the forward run.

The location cusum checks whether the mean of the recursive residuals moves systematically from zero at any point in the sequence. It is

particularly effective in diagnosing changes of regime, omitted predictors, outliers, and some forms of nonlinearity.

The scale cusum checks whether the variance of the recursive residuals in the latter part of the data is different from that in the earlier part. This cusum can be a good indicator of not only heteroscedasticity, but also changes of regime, omitted predictors and nonlinearity.

The test data sets

The use of these diagnostics will now be illustrated with reference to four data sets with various departures from model. All the sets use the same design matrix X with 100 points and 4 predictors. Of the 100 cases, the first 14 and the last 20 are of high leverage (though with the first group on different canonical axes from the second) and the center 66 are of low leverage. The data sets differ in their Y, which contain different departures from model. The individual sets and their departures are as follows:-

1 A set with 10 masked outliers on cases 1 to 10, 4 swamped inliers on cases 11 to 14, and data conforming to model on cases 15 to 100. This is a derivative of the data set used in Hawkins Bradu and Kass (1984) to illustrate the problems of multiple high leverage outliers in regression. Despite the displacement of the outliers by a very large 10 standard deviations, most methods have trouble in detecting them.

2 A set with outliers on the last 20 cases, each point being displaced 3 standard deviations from the generating regression line. While not as difficult to analyse as the first data set using standard methods is not completely transparent.

3 A set with heteroscedasticity, the true residuals on the final 20 points having a standard deviation double that of the first 80 points.

4 A set with a change of regime. The intercept changes by 1 standard deviation after case 80.

With the exception of the first (which was constructed as a counter example to illustrate some deficiencies of current methods), all data sets show departures from model of a type and magnitude that one would hope to be able to detect reliably with standard diagnostics.

The analyses carried out

For each of the data sets we fitted the regression, and computed the ordinary residuals, the recursive residuals computed backwards, and the recursive residuals computed forwards. As a baseline of standard technology using OLS residuals, we analysed the OLS residuals using a normal probability plot and conventional cusums for location and scale (the latter using the square root scale cusum) plugging in the residual standard deviation.

For the recursive residuals we used a normal probability plot and self-starting cusums for location and scale. The computation of the OLS and recursive residuals, and the probability plots of each, were performed using the REGPAC regression program package (Galpin 1981), one of the few general packages available that currently provides recursive residual diagnostics. The self-starting cusums were produced by a FORTRAN subroutine based on that of Hawkins (1987).

As a matter of presentation, all the cusums were drawn in decision

interval form with an allowance per observation of 0.25 standard deviations. With this choice of the allowance, appropriate action limits are horizontal lines at ± 6 standard deviations.

The adequacy of the normal distribution can be tested by the correlation coefficient obtained from the probability plot. The tables of Filliben (1975) show that a correlation coefficient below 0.987 indicates a poor fit at the 5% level of significance for series of length approximately 100.

The plots are given in the appendix. The major features data set by data set are as follows:-

Set 1. The OLS residuals show that there is a problem in this data set. The probability plot rejects normality but while it rightly casts some suspicion on cases 1-10, but wrongly indicates large outliers on cases 11-14 where the largest residuals occur. The conventional cusum for location, but much more so that for scale, show how the overall mean residual and spread drop after case 14 and that the mean rises at the end of the sequence. Most of this diagnosis is incorrect however in that cases 11-100 are good, and only cases 1-10 are outlying.

The backward recursive residuals are also (and not surprisingly) not particularly enlightening as the outliers are never deleted. They show very large values at observations 15 - 17 (the right tail on the probability plot), while the cusums really show little more than that once the high leverage case at the right end of the data set are removed the standard deviations of the recursive residuals get steadily smaller.

The forward cusums provide the correct diagnosis. The probability plot shows how far observations 1-10 deviate (and to a lesser extent also observation 14). The cusums, having run under control up to observation 11, give a very sharp location and scale signal, showing clearly that

observations 1-10 are outliers.

All three probability plots show highly significant non-normality.

Set 2. The OLS residuals show no structure in their probability plot and the correlation coefficient of 0.995 does not correspond to a significant lack of fit. The reason for this is that the outliers are at points of high leverage, an area in which OLS residuals can be expected to perform poorly. The conventional cusums of the OLS residuals do a much better job of detection that there is some problem. The location cusum 'sawtooths' after case 80, giving two upward and two downward signals (a result of the outliers) and the scale cusums move up strongly after case 80, where the outliers start. The diagnosis of the problem however is likely to be completely misleading, as these essentially correct diagnostics are visually overshadowed by the strong downward movement of the scale cusum over most of the sequence which would imply a larger variance on the first few cases than is seen in the rest of the data. This apparent heteroscedasticity is partly real, but primarily an artefact due to the smearing by OLS of the effect of the departure in the right end of the data, the misfit being magnified in the left end of the data by the high leverage of these cases.

The probability plot of the backward recursive residuals indicates significant non-normality as one would hope of a data set with twenty 3σ outliers, though the plot does not clearly show the problem as outliers. The corresponding location and scale cusums however provide significant moves starting at observation 80, for an essentially correct diagnosis. We note that there is also a signal of decreased variance starting about case 10. This reflects the fact that among the first 10 recursive residuals, by chance three exceed 2σ and the running standard deviation at 10 observations is 1.53σ. Thus the signal correctly reflects a changing variance, though the reason for the change is the random numbers obtained in the simulation rather than a model defect.

As the departure from model is at the right end of the data set, one should not expect to see the correct diagnosis in the forward recursive residuals, but in fact these do provide a correct diagnosis. There is a significant location signal with the cusum peaking at observation 20, but the outliers' real footprint is the steady reduction in the variance past observation 21.

The probability plot shows significant non-normality and is perhaps somewhat clearer than its backward counterpart in showing two clearly separated outliers.

Set 3. The probability plot of the OLS residuals is unremarkable. It has apparent gaps at the left and right, but these are not so large as to give a significant departure from model. While the cusums of the OLS residuals provide a number of significant signals in both location and scale, the correct diagnosis of an increased variance after case 80 is not seen.

The probability plot of the backward recursive residuals shows no departure from model, but the cusum apart from a weak and nonpersistent signal of decreased variance in the middle of the sequence, clearly shows a variance increase starting at observation 79 - essentially the correct conclusion.

The probability plot of the forward recursive residuals shows just-significant non-normality, with an excess of large negative residuals apparent. The cusum amplifies this diagnosis by discovering that the spread of the residuals decreases significantly once the earlier points are added to the later ones. The variance change point is diagnosed as being between cases 78 and 79, as with the backward cusums and again providing the correct diagnosis.

Set 4. The probability plot of the OLS residuals shows no significant non-normality, and looks acceptable apart from an apparent outlier on the left (which is in fact a good observation). The cusum shows some departure from model, but the plethora of signals gives no clear indication of what it may be - at various points in the sequence, both upward and downward signals are seen in both the location and the scale cusums.

The probability plot of the backward recursive residuals is also close to linear. The location cusum localises the problem correctly however, giving a significant upward shift at case 87. There is a false alarm in the form of a significant downward scale signal for portions of the middle of the set, but these signals are much weaker than that of the location shift.

The forward recursive residuals show the same apparent outlier seen in the OLS residuals. They give an acceptably linear probability plot, but one displaced from the origin - a feature on whose importance we commented above and which is indicative of problems in the cases that are never deleted. The location cusum shows a significant downward shift between cases 88 and 89 - a correct diagnosis of the departure, though not at quite the correct position in the sequence.

Summary and conclusions

All four of the data sets had one feature in common:- a departure from model that affected a portion of the data but left one end of the sequence fitting the model. It is under these circumstances that the approach of using recursive residuals, along with the self-starting cusums which capitalise on their particular properties, can be expected to be most effective. In all four cases, it was shown how these methods

provided a much clearer diagnosis of the model defect than was obtainable from the more traditional analyses using OLS residuals and conventional cusums.

We pointed out in the discussion of the effect of model departures the importance of their being an ordering such that at least one of the two ends of the data set consisted only of data for which the model is appropriate. This situation will not always apply, and the deficiency may only show up with some other ordering of the data.

By implication, we have been talking about the data as processed in lexicographic order. If however we were interested to check the linearity of the regression on a particular one of the predictors, then it would be necessary to reorder the data on that predictor and compute the forward and backward cusum diagnostics on that ordering. There are many such potential orderings that may legitimately be used - in particular one could envisage routine use each of the predictors and/or the predicted Y as the basis for ordering and the computation of the recursive residuals. Another interesting possibility (suggested to the author by Alan Dorfman) is ordering by increasing values of the leverage, with deletion from the end of high leverage. This latter possibility should be particularly effective in detecting smooth nonlinearities, as it would consist essentially of testing the points with remote x values against those with central values and so giving improved chances of locating 'the outliers that matter'.

This large number of potential plots does not involve as much computation as one might suppose - any reordering is an O(nlogn) operation, and the computation of a set of recursive residuals an $O(np^2+p^3)$ operation. The major obstacle to making all p+3 orderings is not the computational load, but the visual one of scanning all the resultant outputs. Here the recommendation is of reporting by exception; of programming so that only those cusums giving significant

signals are displayed for closer interpretation. There are clear 'cognostic indices' of which plots to inspect in the form of the correlation coefficient of the probability plot, a t statistic for overall non-zero mean on the recursive residuals, and the maxima and minima of the location and scale cusums. These indices may be used to rank the possible orderings from most to least interesting, and since they correspond to formal significance tests, may also be used to filter out plots in which the model gives no significant misfit.

A departure from model which does not affect one end of the data is not the universal norm, and we do not claim that the probability plot / self-starting cusum diagnostic is a panacea for all model diagnoses; nevertheless we consider that the situation modelled in the test data sets and the extension to other orderings is common enough to argue for the availability of these procedures in all regression packages.

References

Brown, R. L.,, Durbin, J., and Evans, J. M., (1975), 'Techniques for testing the constancy of regression relationships over time' (with Discussion), Journal of the Royal Statistical Society, 37, 149 - 192.

Filliben, J. J., 'The probability plot correlation coefficient test for normality', Technometrics, 17, 111 - 117.

Galpin, J. S., (1981), 'Regression Package REGPAC' Report SWISK 25, Council for Scientific and Industrial Research, Pretoria.

Galpin, J. S., and Hawkins, D. M., (1984), 'The use of recursive residuals in checking model fit in linear regression', The American Statistician, 38, 94 - 105.

Garbade, K., (1977), 'Two methods for examining the stability of regression coefficients', Journal of the American Statistical Association, 72, 54 - 63.

Gragg, W. B., LeVeque, R. J., and Trangenstein, J. A., (1979), 'Numerically stable methods for updating regressions', Journal of the American Statistical Association, 74, 161-168.

Hawkins, D. M., (1981), 'A cusum for a scale parameter' Journal of Quality Technology', 13, 228 - 231.

Hawkins, D. M., (1987), 'Self-starting cusum charts for location and scale', to appear in The Statistician.

Hawkins, D. M., Bradu, D., and Kass, G. V., (1984), 'Location of several outliers in multiple regression data using elemental sets', Technometrics, 26, 197 - 208.

Hester, R. A., and Quesenberry, C. P., (1984), 'Analyzing uniform residuals for heteroscedasticity', Institute of Statistics Mimeo Series No 1639, North Carolina State University, Raleigh, NC.

Maindonald, J. H., (1984), 'Statistical Computation', Wiley, New York.

McCabe, B. P. M., and Harrison, M. J., (1980), 'Testing the constancy of regression relationships over time using least squares residuals', Applied Statistics, 29, 142 - 148.

O'Reilly, F. J., and Quesenberry, C. P., (1973), 'The conditional probability integral transformation and applications to obtain composite chi-square goodness of fit tests', Annals of Statistics, 1, 74-83.

Peizer, D. B., and Pratt, J. W., (1968), 'A normal approximation for

binomial, F, beta and other common related tail probabilities I' Journal of the American Statistical Association, 63, 1416 - 1456.

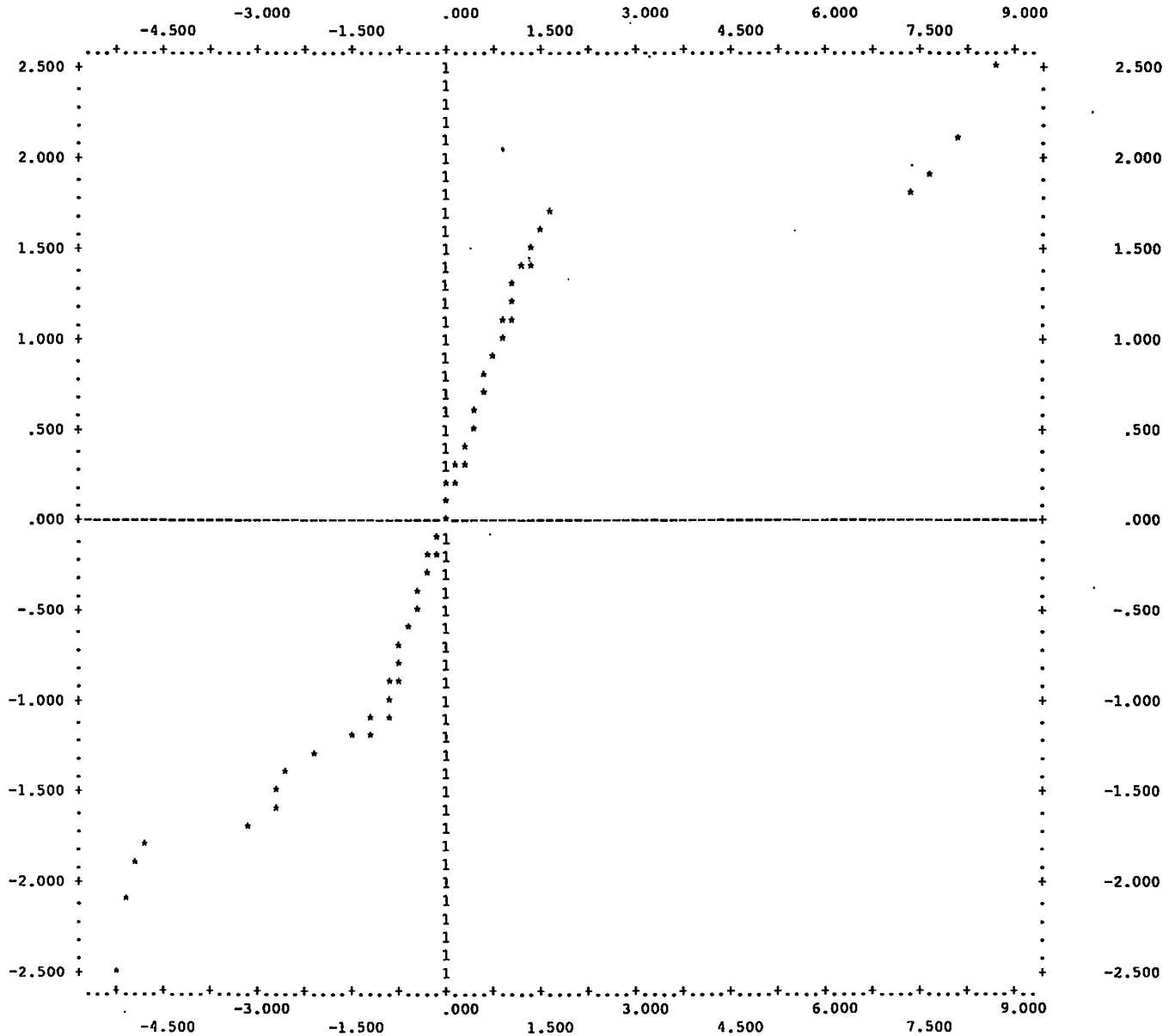
Legend to figures

The cusums are of the decision interval form with an allowance of 0.25 standard deviations per observation. The action limits are at ± 6. Three cusums are given for each data set - a conventional cusum for the OLS residuals, and a self-starting cusum for the backward, and for the forward recursive residuals.

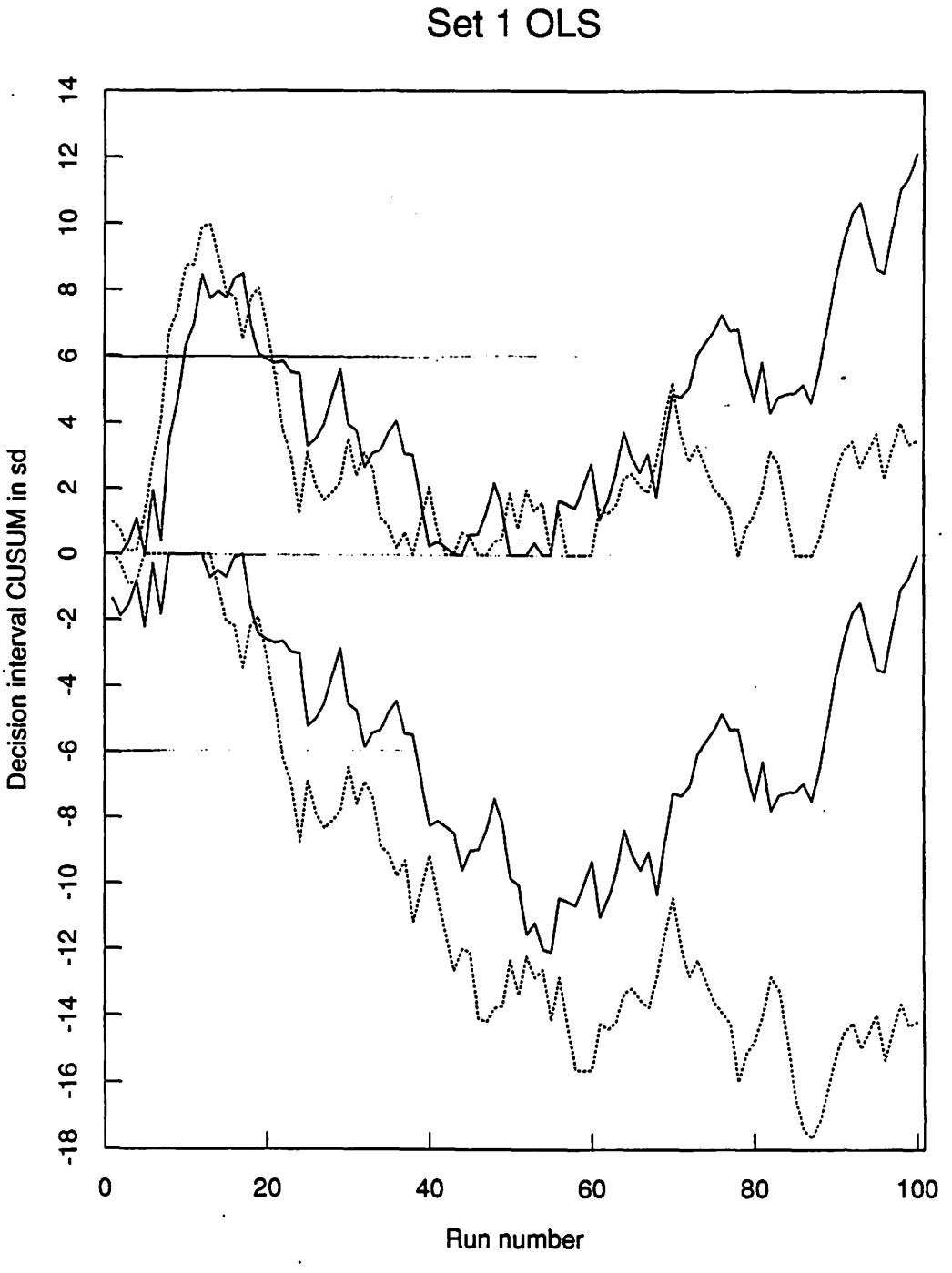
The location cusums are shown as solid lines and the scale cusums as dotted. The positive half shows increases, and the negative half decreases in location and scale respectively.

Note that in the cusums of the OLS and the backwards recursive residuals, the points are plotted in the same order as they occur in the data set, but for the forward recursive residuals they are reversed - the point with index 1 is the last point in the data set and not the first.

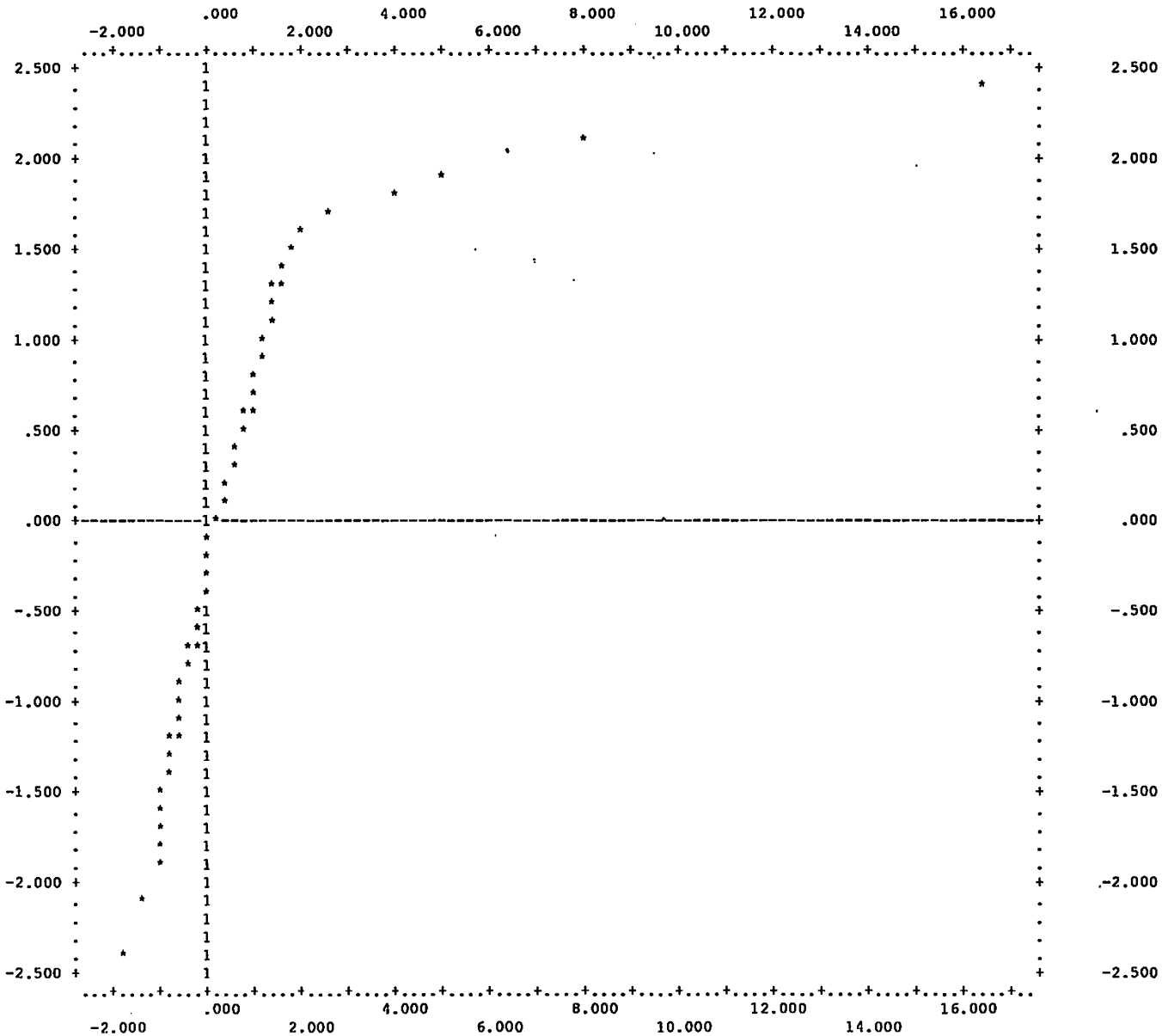
Set 1. Probability plot of OLS residuals



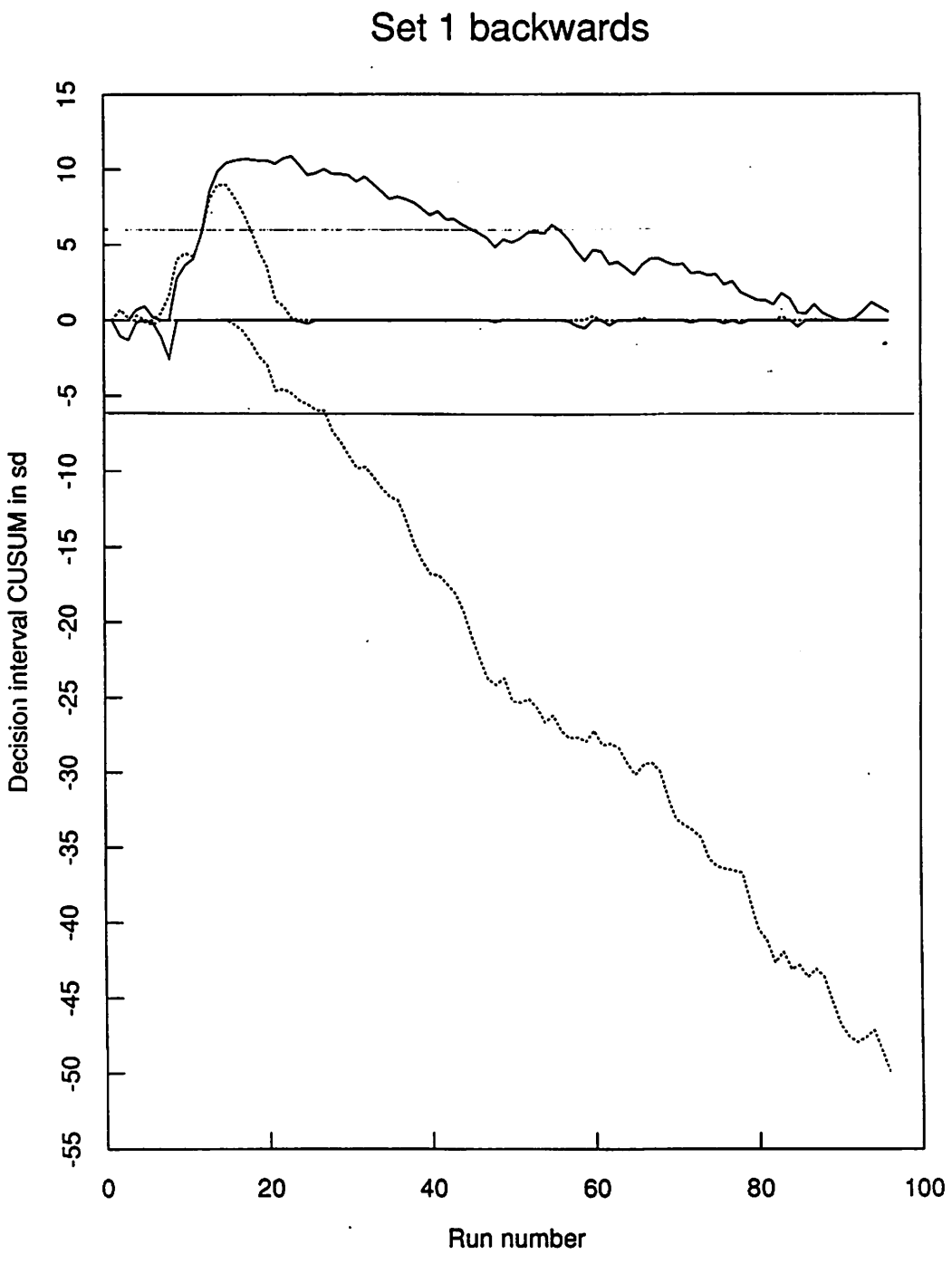
NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .859979


Set 1 OLS
Decision interval CUSUM in sd
14
12
10
8
6
4
2
0
-2
-4
-6
-8
-10
-12
-14
-16
-18
0
20
40
60
80
100
Run number

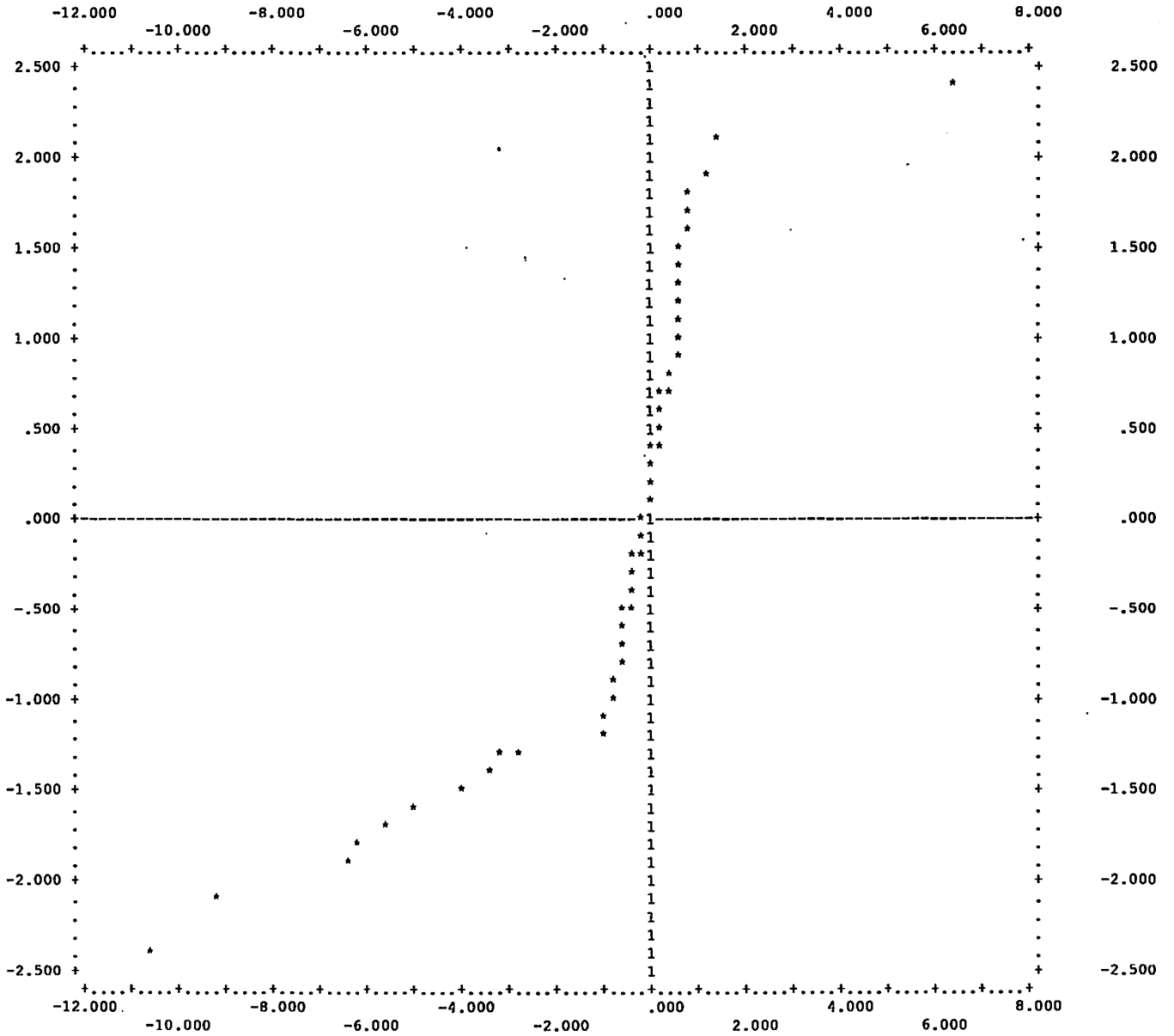
Set 1. Probability plot of backward recursive residuals



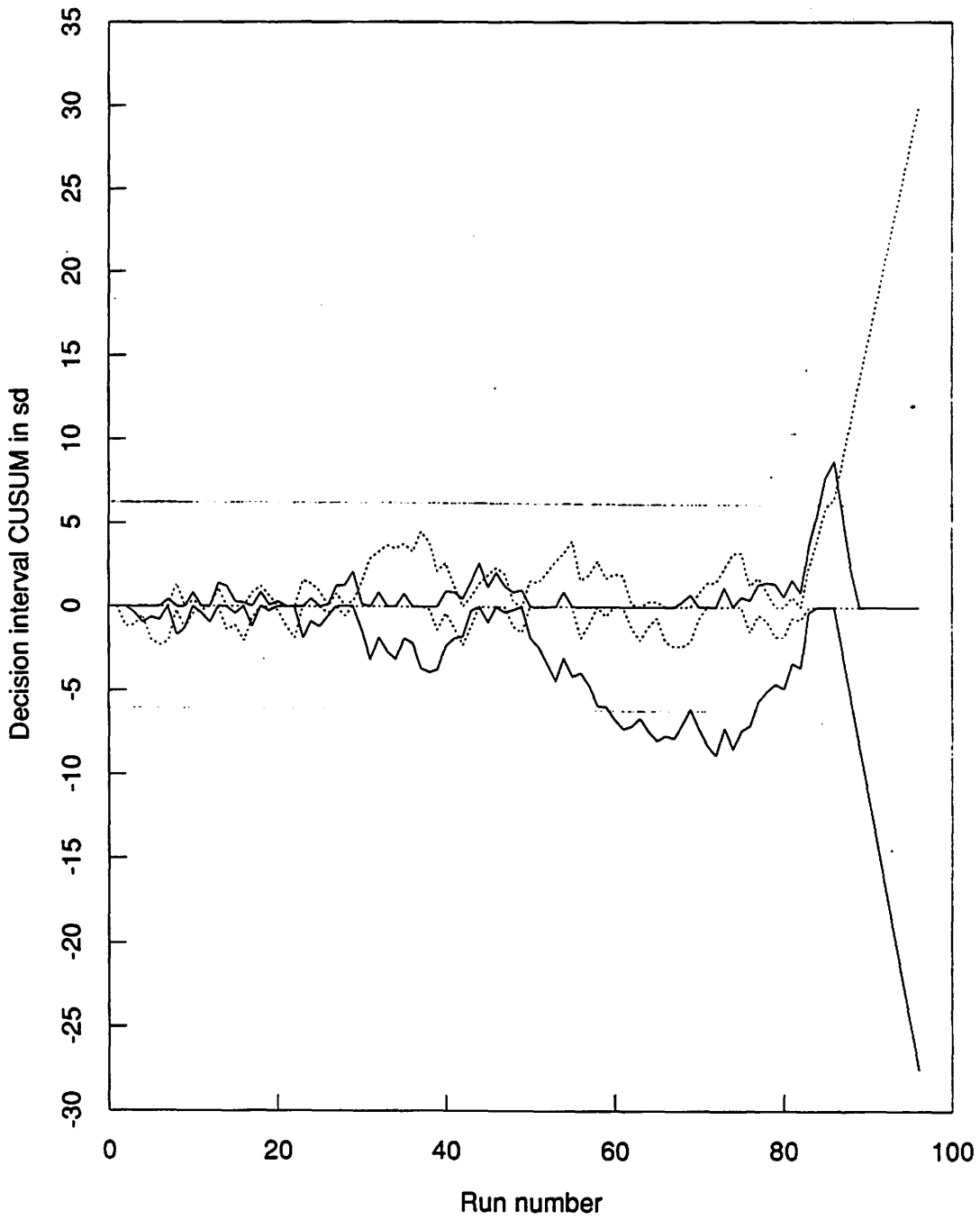
NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .716153


Set 1 backwards
Decision interval CUSUM in sd
15
10
5
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
-50
-55
0
20
40
60
80
100
Run number

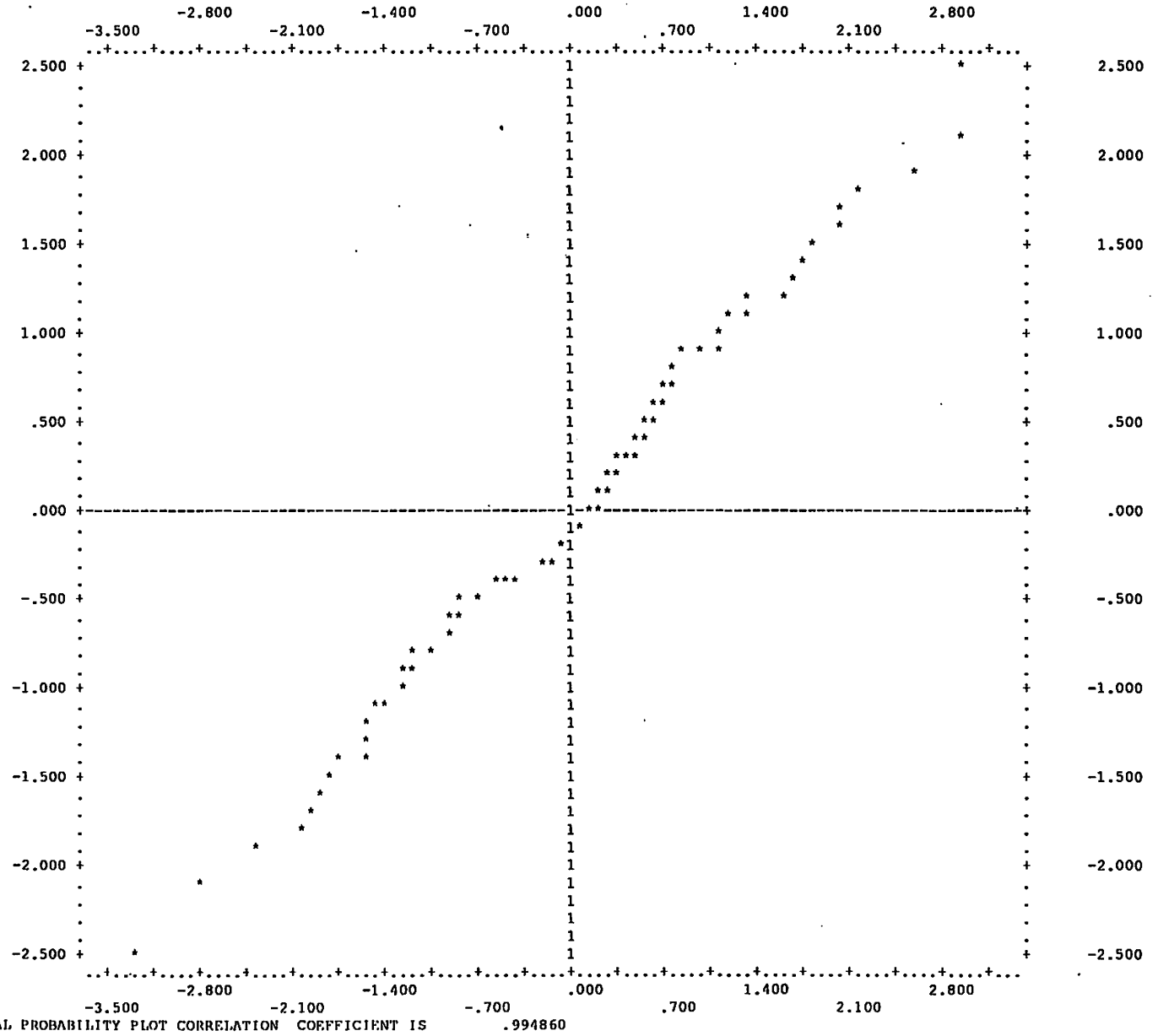
Set 1. Probability plot of forward recursive residuals



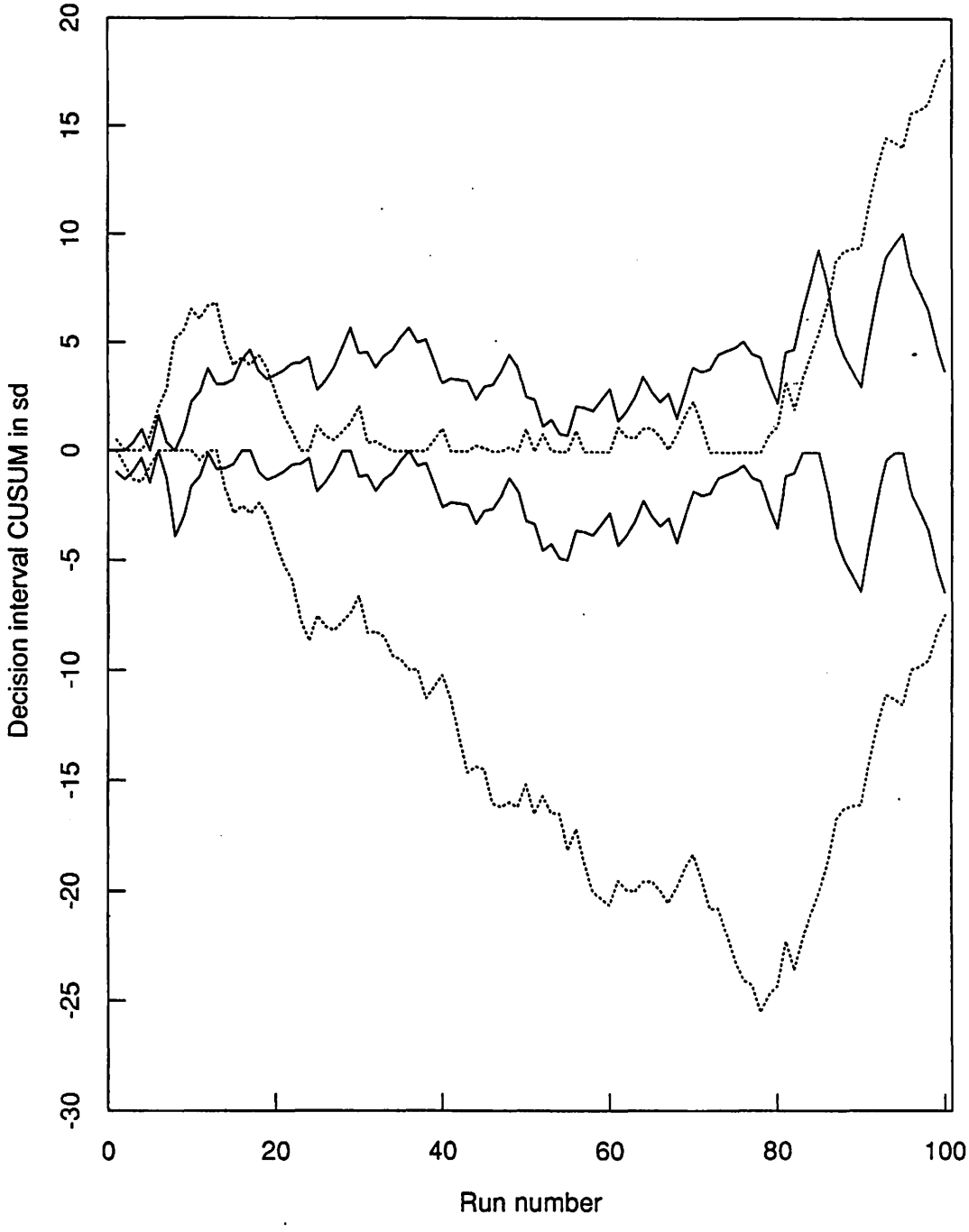
NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .791589


Set 1 forwards
Decision interval CUSUM in sd
35
30
25
20
15
10
5
0
-5
-10
-15
-20
-25
-30
0
20
40
60
80
100
Run number

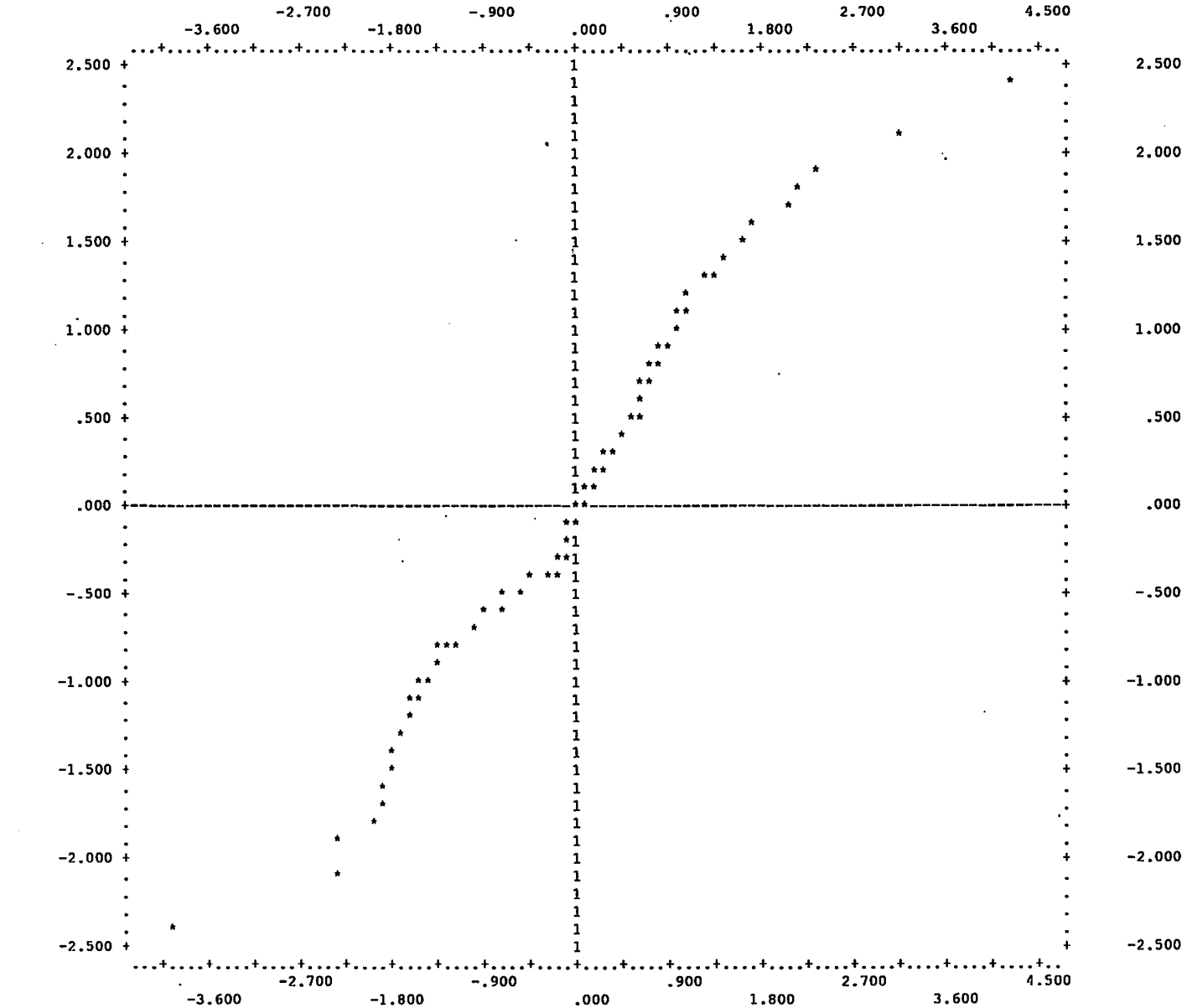
Set 2. Probability plot of OLS residuals



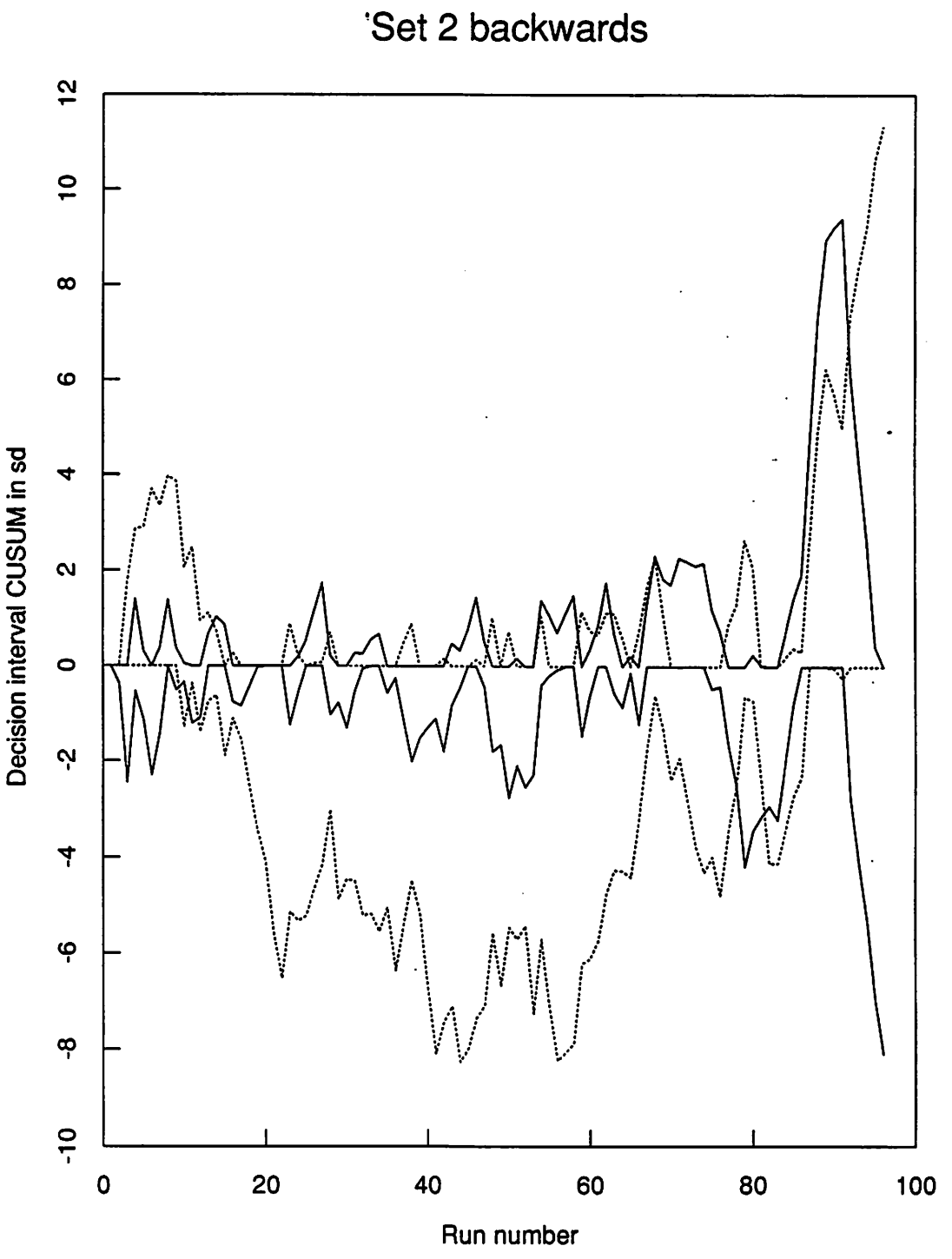
NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .994860


Set 2 OLS
Decision interval CUSUM in sd
20
15
10
5
0
-5
-10
-15
-20
-25
-30
0
20
40
60
80
100
Run number

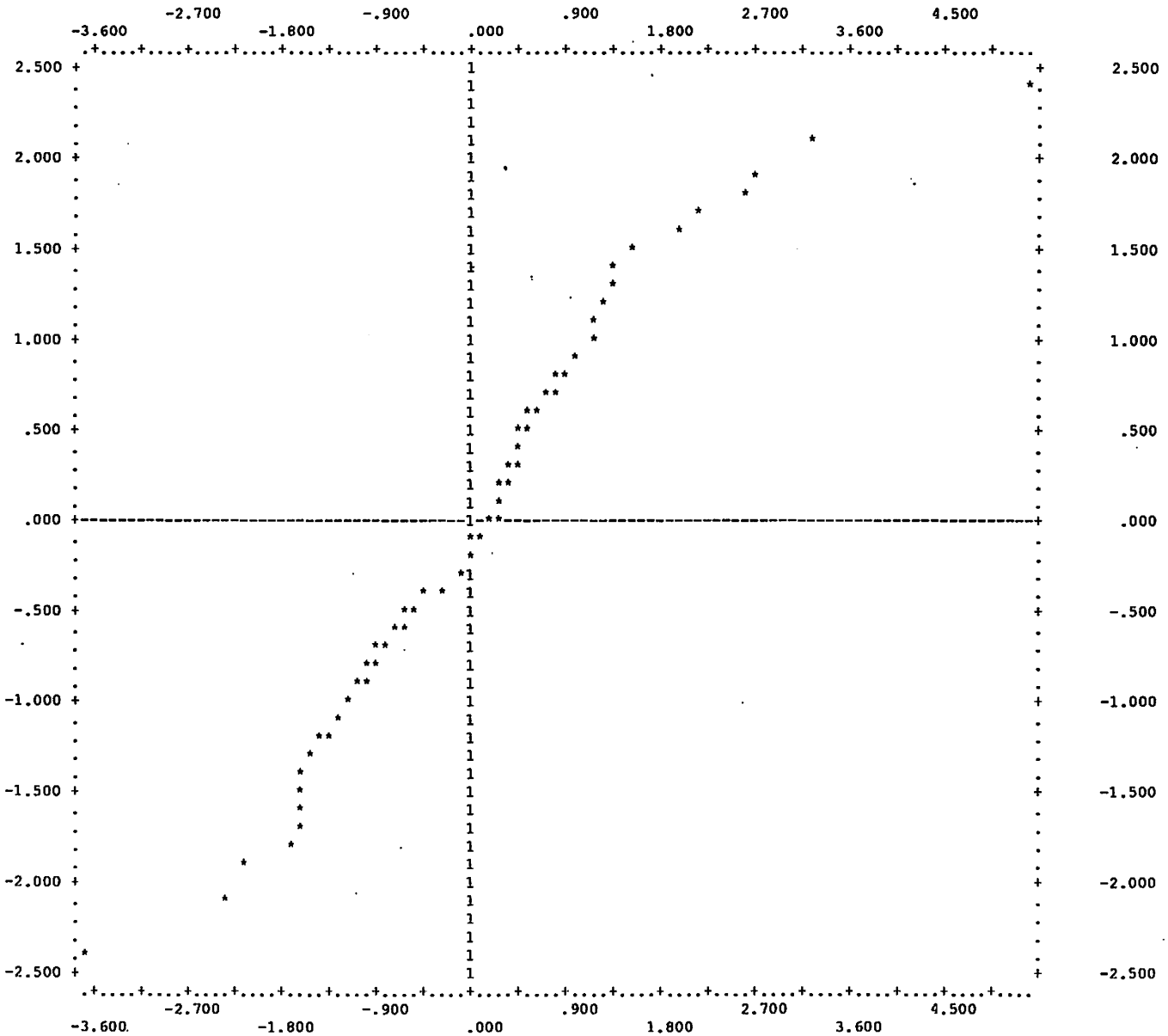
Set 2. Probability plot of backward recursive residuals



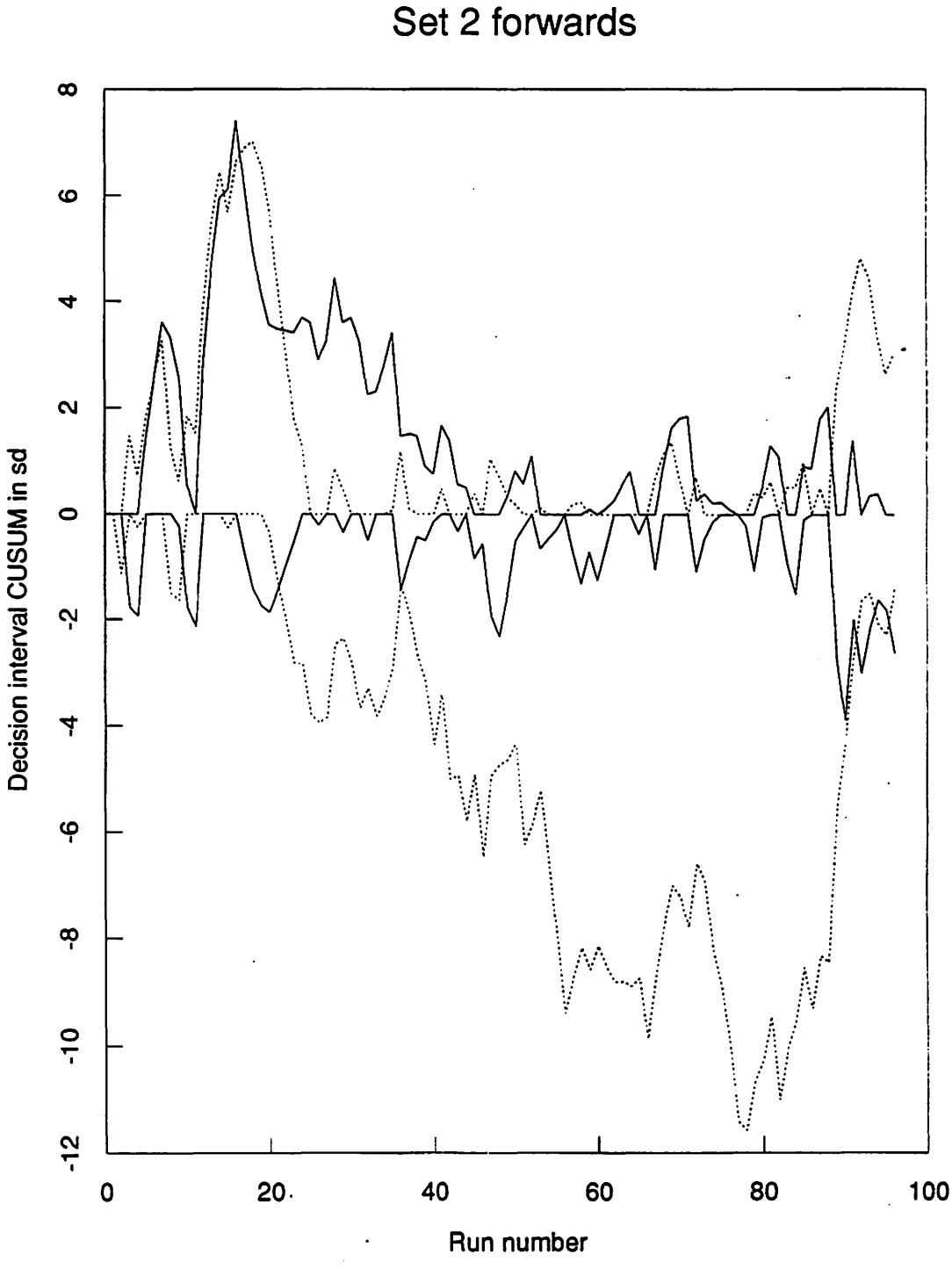
NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .982858


'Set 2 backwards
Decision interval CUSUM in sd
12
10
8
6
4
2
0
-2
-4
-6
-8
-10
0
20
40
60
80
100
Run number

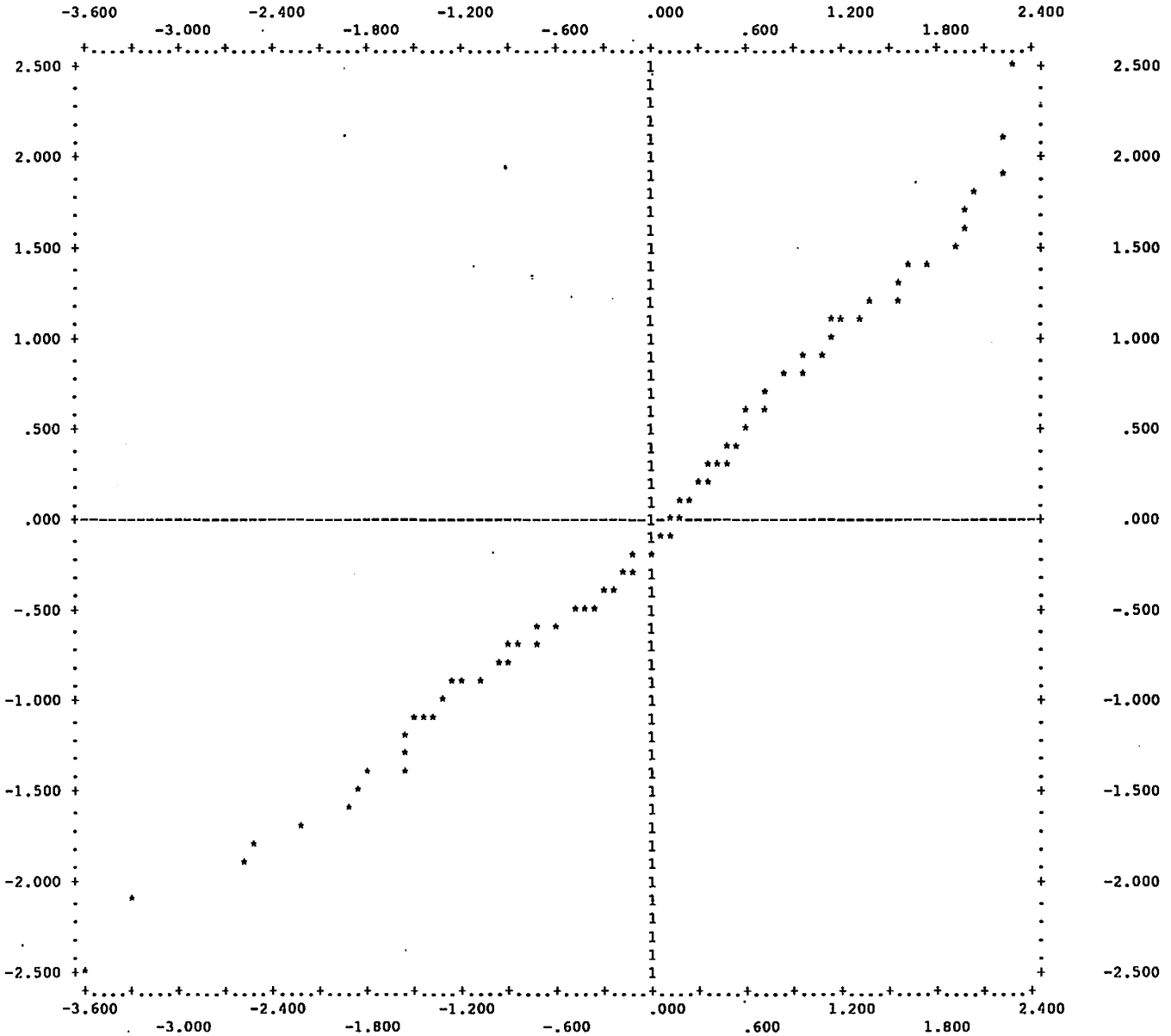
Set 2. Probability plot of forward recursive residuals



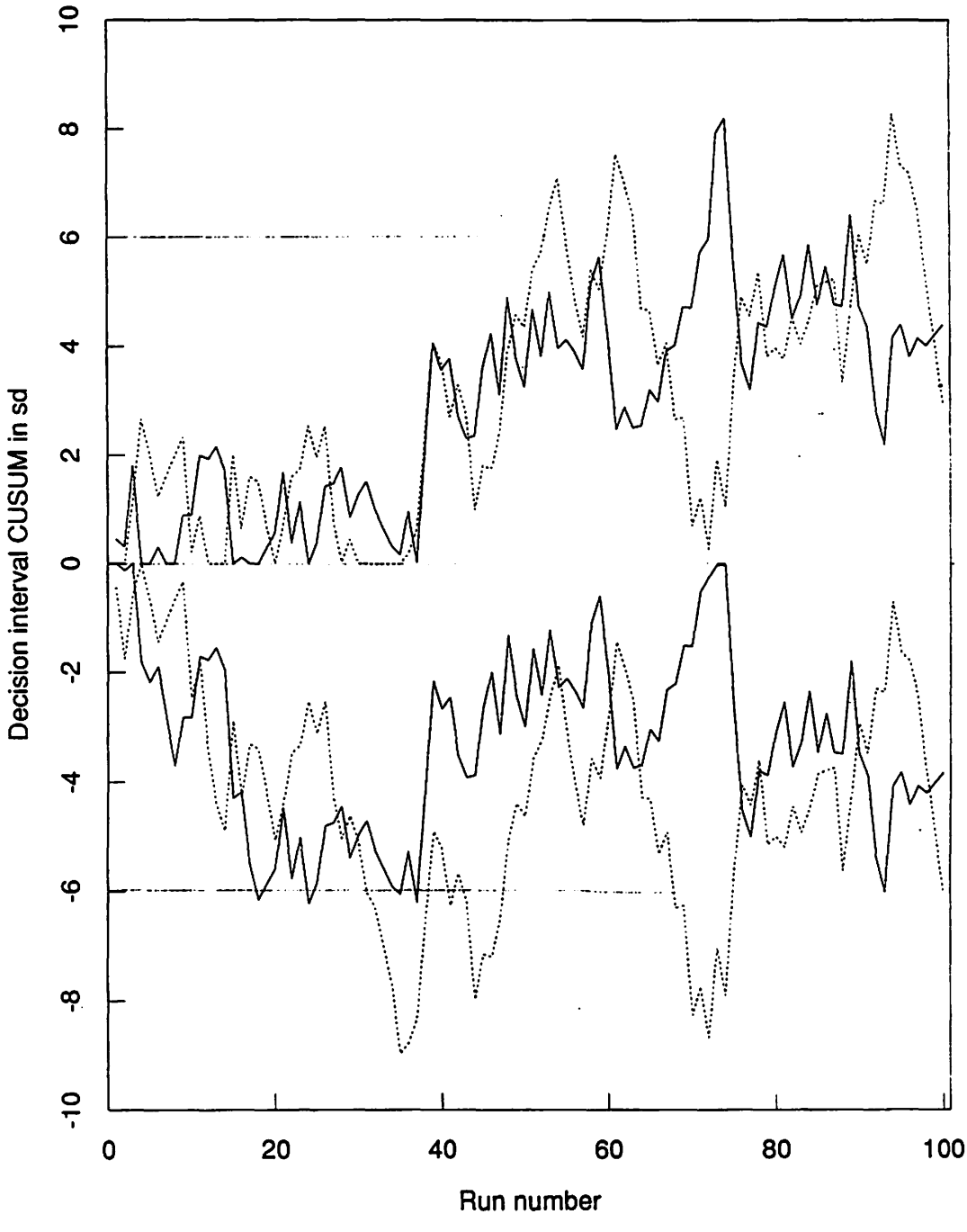
NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .971927


Set 2 forwards
Decision interval CUSUM in sd
8
6
4
2
0
-2
-4
-6
-8
-10
-12
0
20
40
60
80
100
Run number

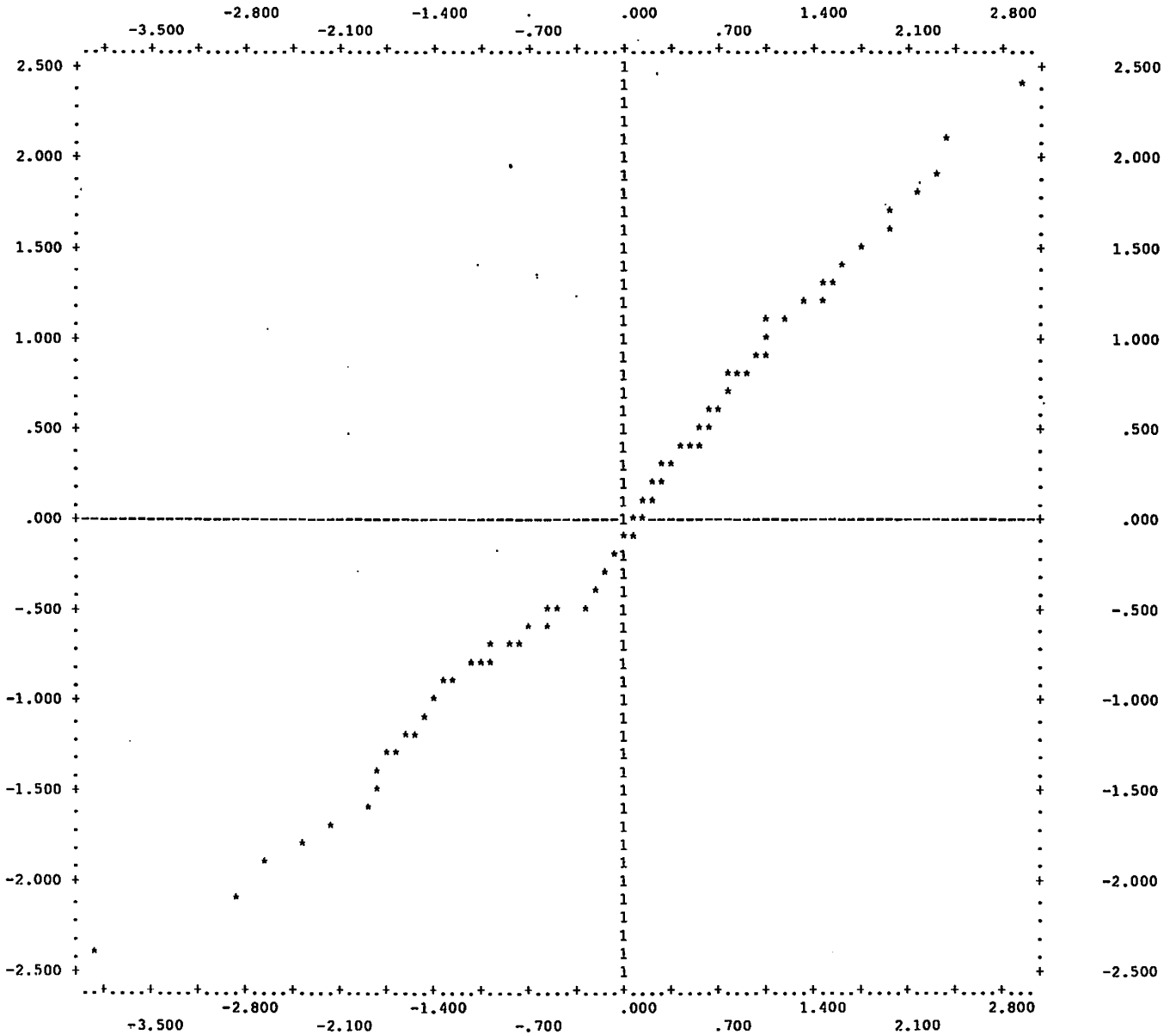
Set 3. Probability plot of OLS residuals



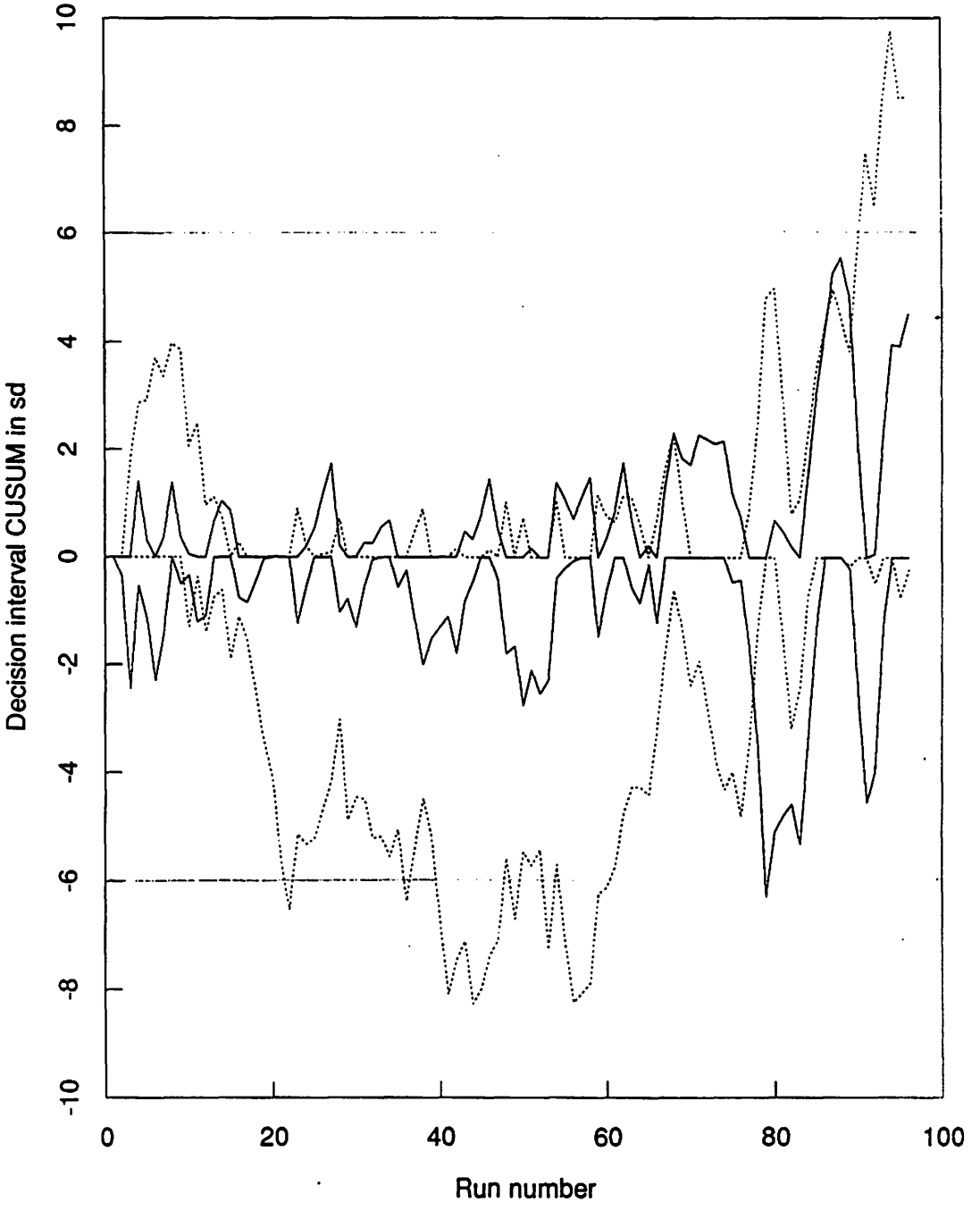
NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .989538


Set 3 OLS
Decision interval CUSUM in sd
10
8
6
4
2
0
-2
-4
-6
-8
-10
0
20
40
60
80
100
Run number

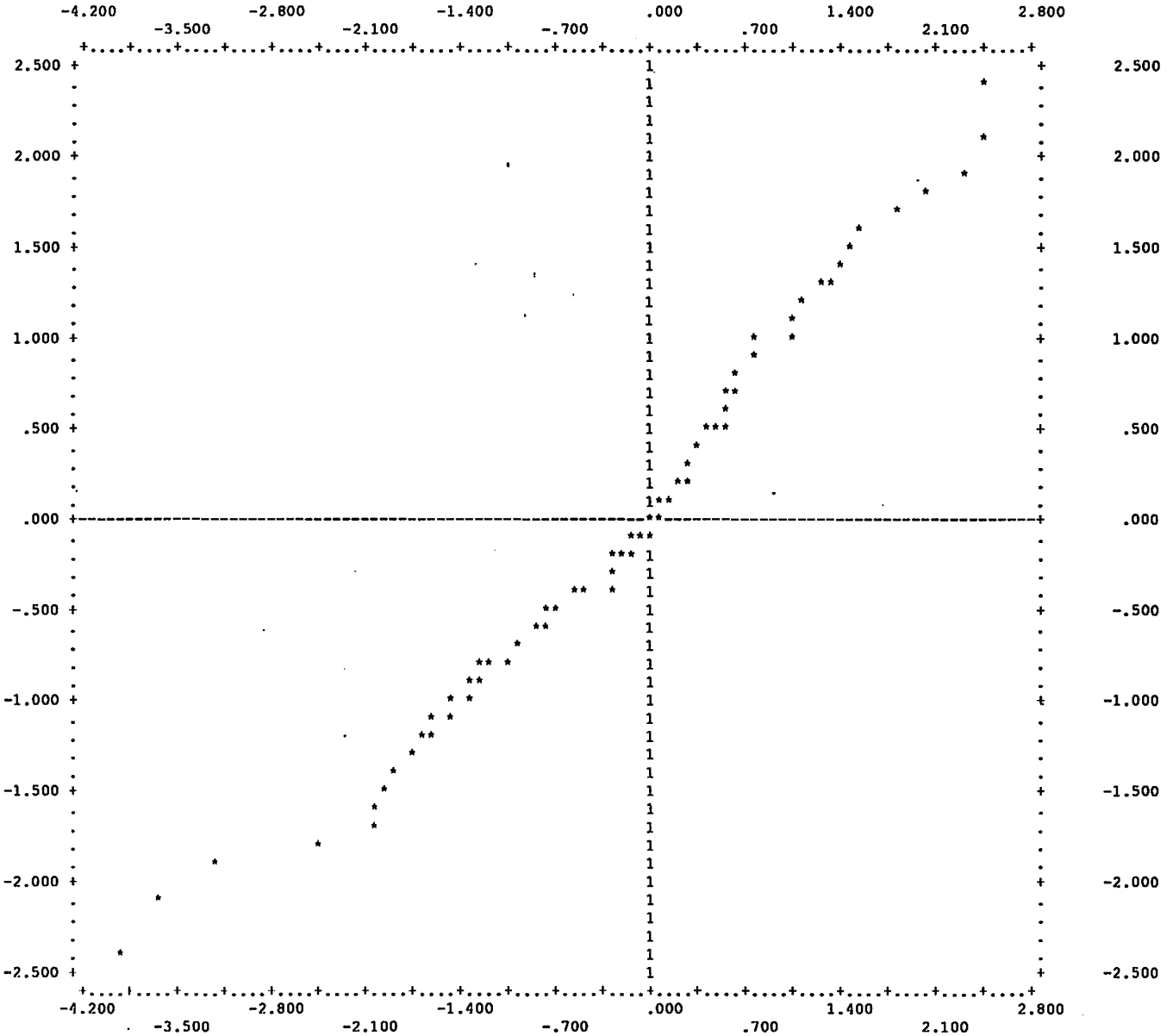
Set 3. Probability plot of backward recursive residuals



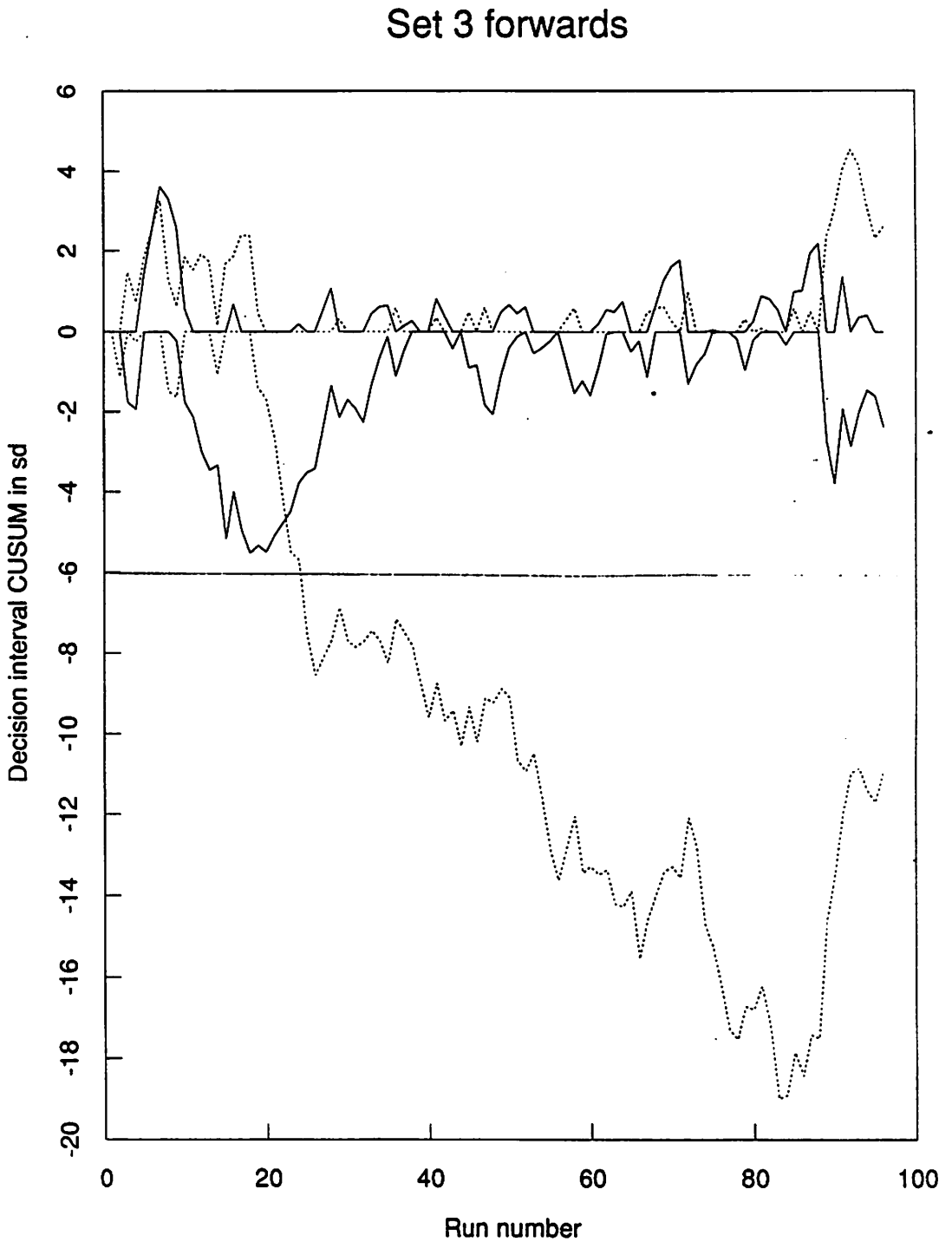
NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .991020


Set 3 backwards
Decision interval CUSUM in sd
10
8
6
4
2
0
-2
-4
-6
-8
-10
0
20
40
60
80
100
Run number

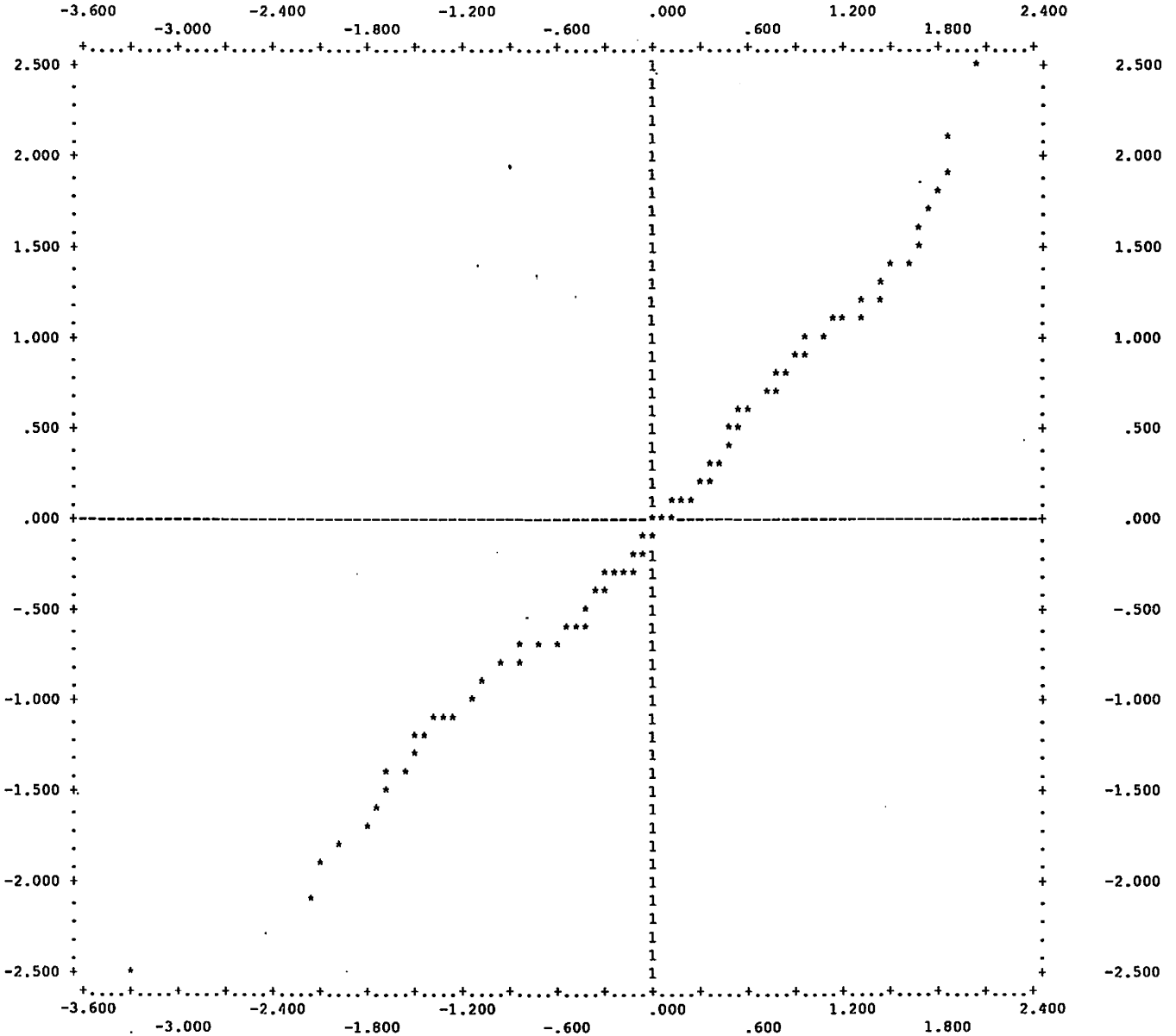
Set 3. Probability plot of forward recursive residuals



NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .986736

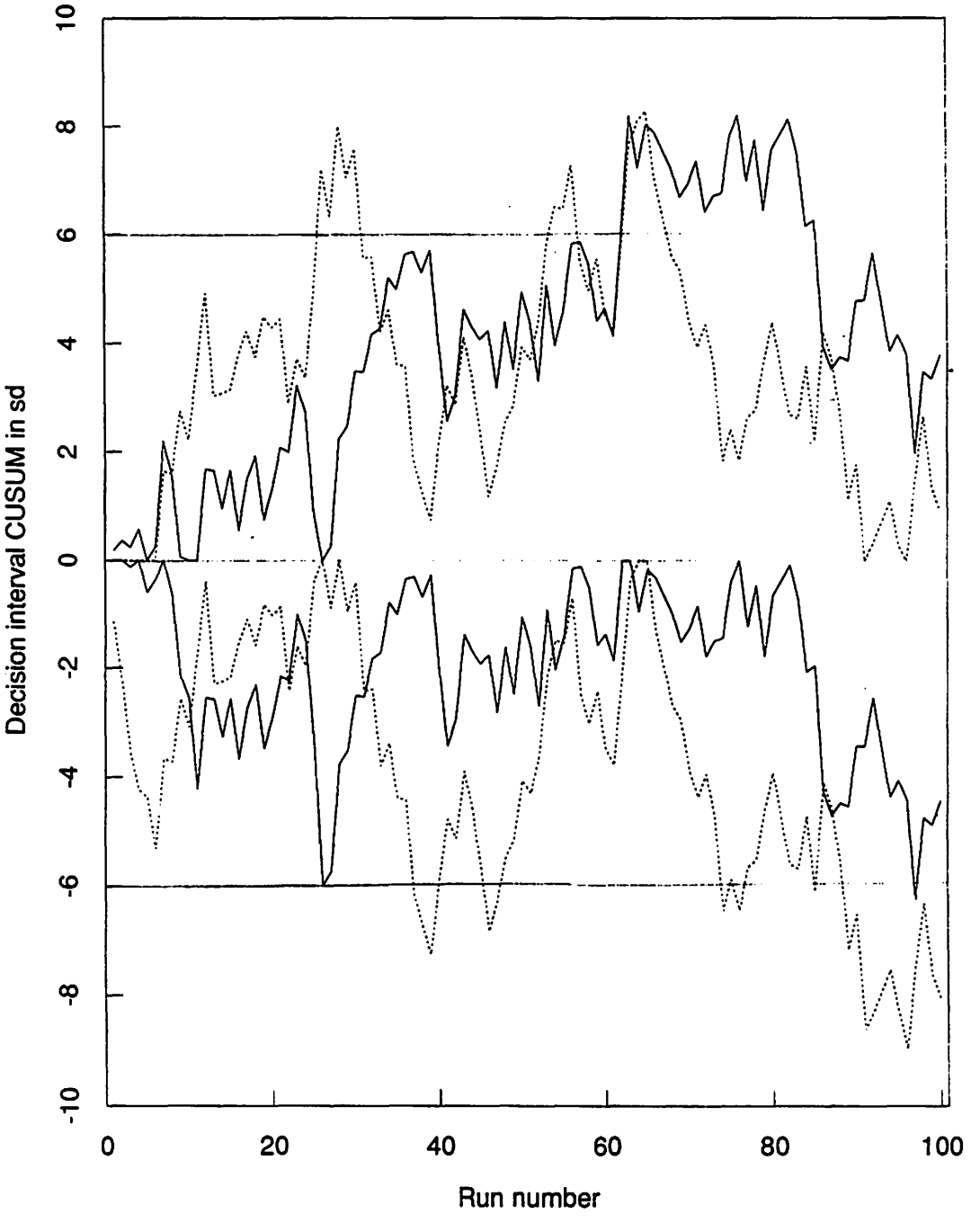

Set 3 forwards
Decision interval CUSUM in sd
6
4
2
0
-2
-4
-6
-8
-10
-12
-14
-16
-18
-20
0
20
40
60
80
100
Run number

Set 4. Probability plot of OLS residuals

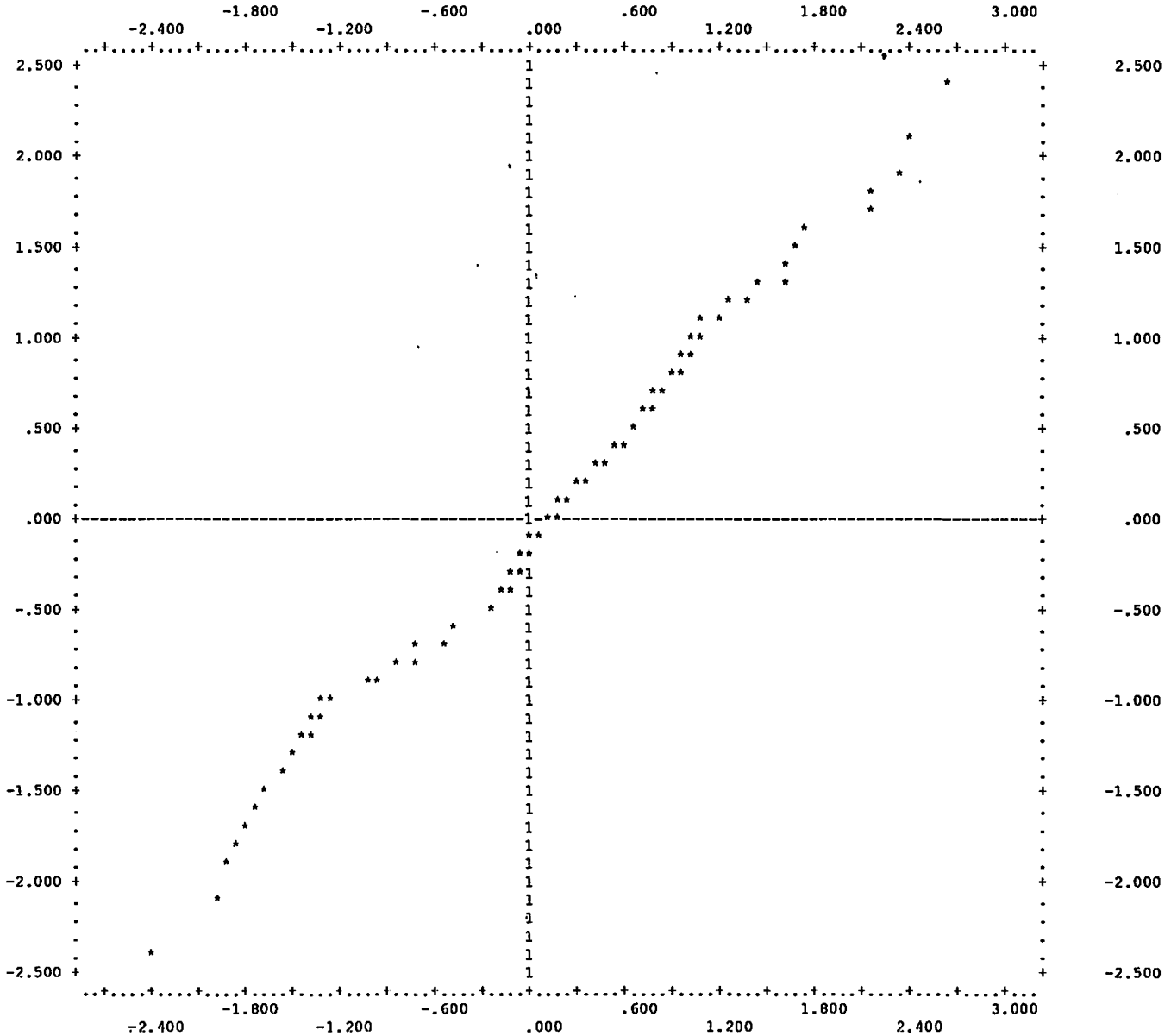


NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .991988

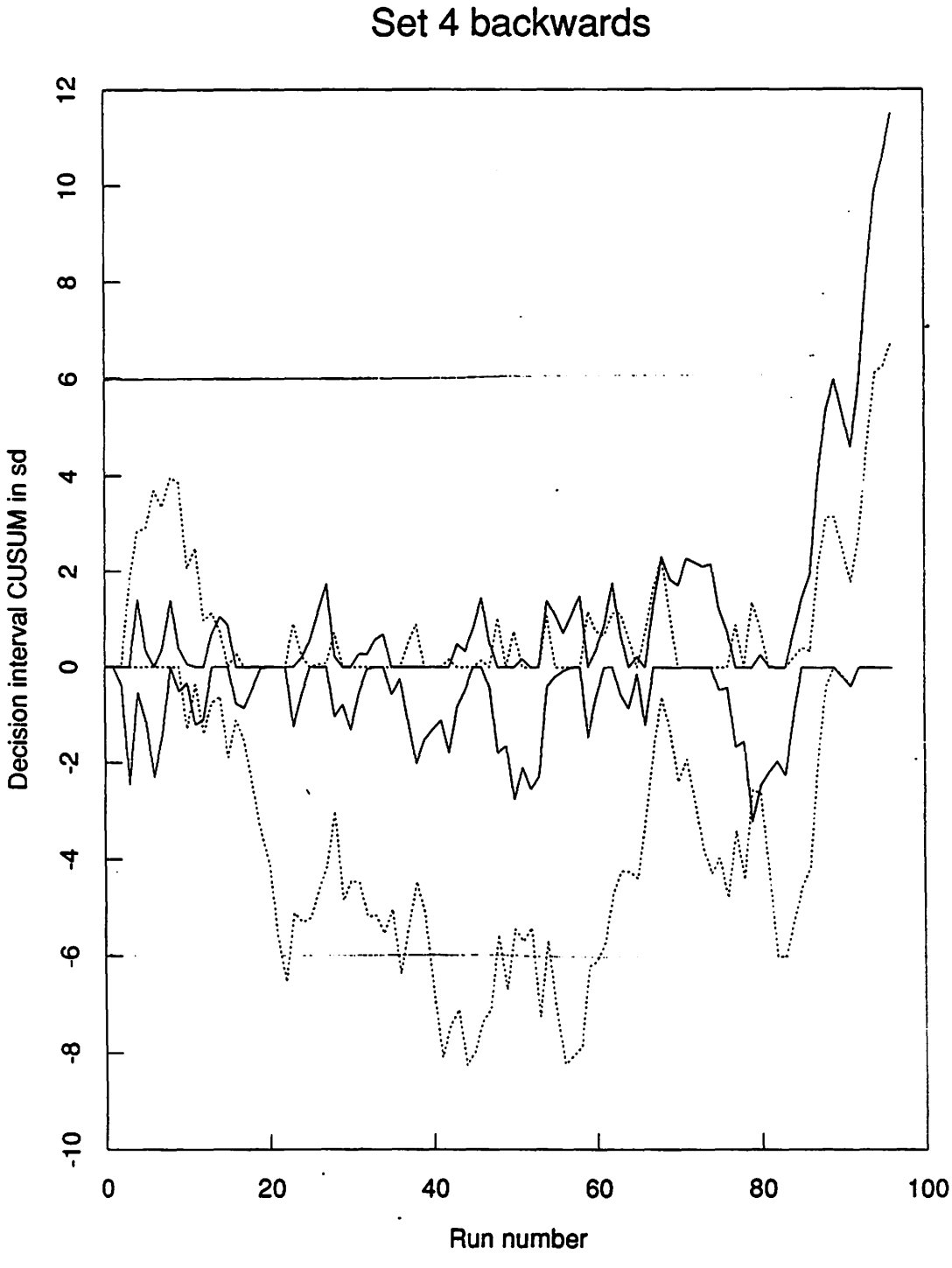
Set 4 OLS


Decision interval CUSUM in sd
10
8
6
4
2
0
-2
-4
-6
-8
-10
0
20
40
60
80
100
Run number

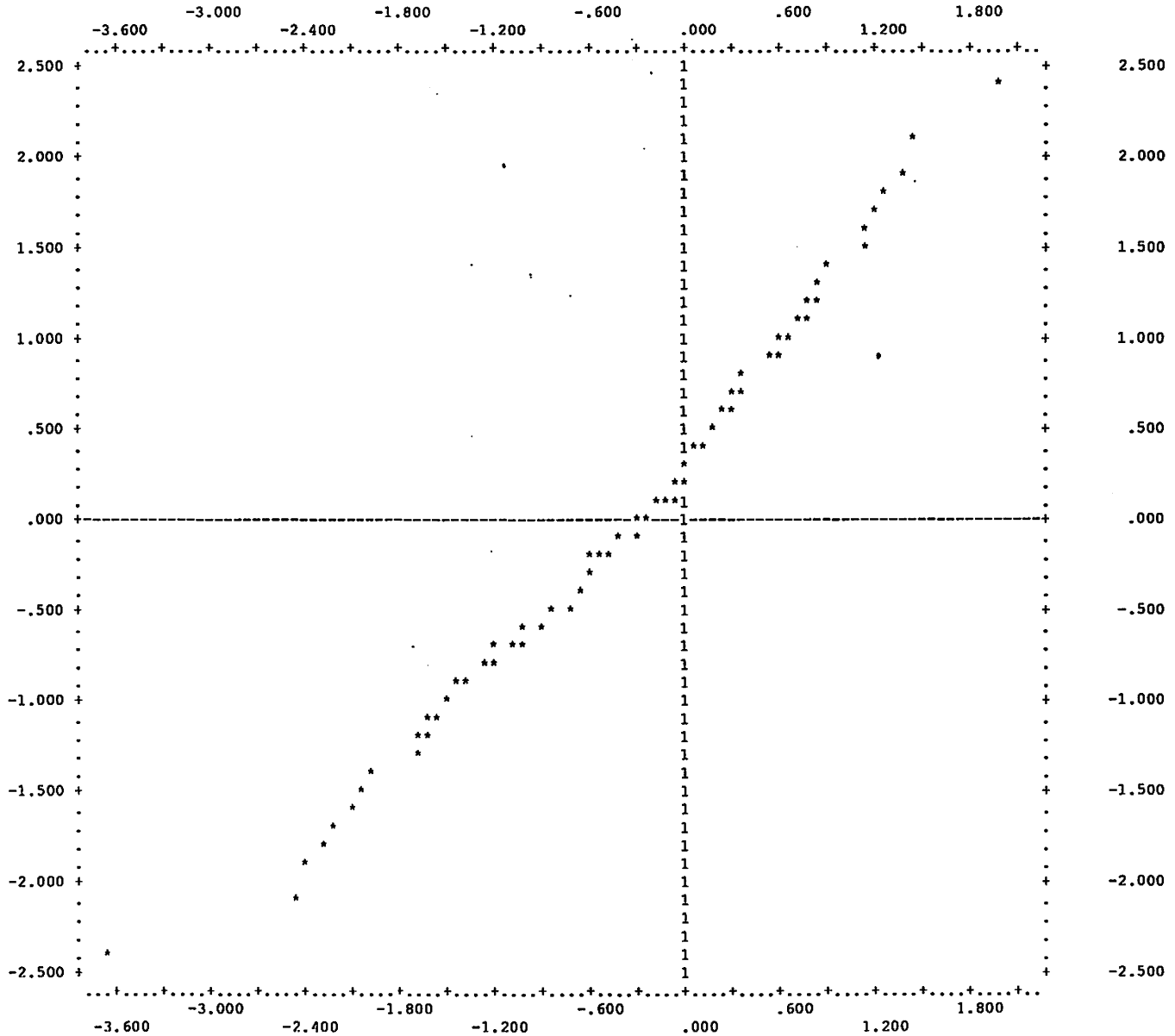
Set 4. Probability plot of backward recursive residuals



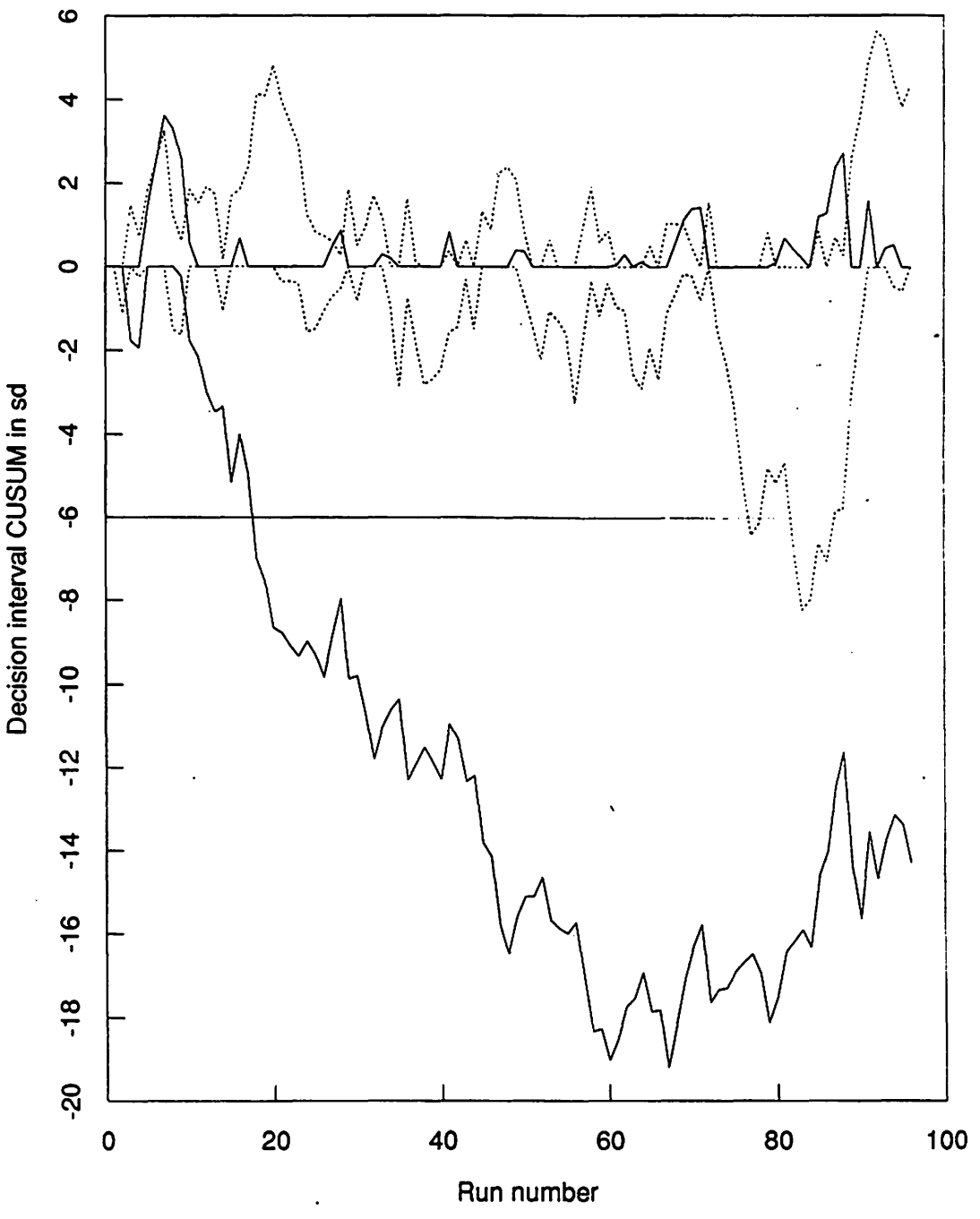
NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .994070


Set 4 backwards
Decision interval CUSUM in sd
12
10
8
6
4
2
0
-2
-4
-6
-8
-10
0
20
40
60
80
100
Run number

Set 4. Probability plot of forward recursive residuals



NORMAL PROBABILITY PLOT CORRELATION COEFFICIENT IS .992409


Set 4 forwards
Decision interval CUSUM in sd
6
4
2
0
-2
-4
-6
-8
-10
-12
-14
-16
-18
-20
0
20
40
60
80
100
Run number