A PROGRAM FOR ESTIMATING THE POINT OF CHANGE AND DEGREE IN POLYNOMIAL REGRESSION

by

S.R. Esterby

NWRI CONTRIBUTION #85-96

National Water Research Institute
Canada Centre for Inland Waters
867 Lakeshore Road, P.O. Box 5050
Burlington, Ontario, Canada L7R 4A6

A PROGRAM FOR ESTIMATING THE POINT OF CHANGE AND DEGREE IN POLYNOMIAL REGRESSION

S.R. Esterby

Aquatic Ecology Division

National Water Research Insitute

Burlington, Ontario, Canada L74 4A6

March 1985

National Water Research Institute, Contribution 85-96

Submitted to IWD Scientific Series

RESUME ADMINISTRATIF

Une méthode a été conçue (Esterby et El-Shaarawi, 1981) pour permettre d'appliquer l'analyse de la régression aux cas où l'état d'un système est perturbé et la relation entre les variables change au cours d'une même période d'observation. Cette méthode a été mise au point surtout pour servir de modèle objectif dans l'évaluation de l'horizon à ambroisies (horizon qui sert de repère temporel dans l'étude des carottes de sédiments lacustres) mais elle peut avoir des applications beaucoup plus vastes dans le domaine des sciences de l'eau. Elle pourrait, par exemple, servir à déterminer à quel moment les paramètres de la qualité de l'eau changent après qu'on ait pratiqué une coupe rase des boisés d'un bassin de drainage ou éliminé une source de pollution. A l'Institut national de recherche sur les eaux, on a largement eu recours à cette méthode pour traiter les données sur le pollen fossile et les diatomées rassemblées dans le cadre du Programme de transport à distance des polluants atmosphériques. Un programme FORTRAN visant à permettre aux intéressés d'appliquer la méthode vient d'être rédigé et ce rapport explique comment l'utiliser. Puisqu'il s'agit d'une nouvelle méthode qui ne figure encore dans aucun manuel didactique, la publication de ce rapport s'impose sil l'on vert que tous les intéressés puissent y avoir accès.

RESUME

On fait la description d'un programme FORTRAN portant sur une méthode d'évaluation d'un point de changement dans un modèle de régression de polynôme. Le modèle se fonde sur les mêmes axiomes que ceux qui s'appliquent à un polynôme unique mais on suppose que le polynôme, et peut-être aussi les variances, ont été modifiés à un point de l'ensemble des données. Il est également important de connaître l'ordre des points de données. On évalue le degré du polynôme de chaque segment et le point de changement en appliquant un procédé itératif qui fait appel à la fonction de probabilité marginale pour déterminer le paramètre point de changement et un test qui utilise les variations dans les sommes des carrés des écarts avec réduction de degré. On peut vérifier les axiomes du modèle grâce aux écarts et à la fonction de probabilité. Des exemples illustrent l'utilité du programme et les résultats sont présentés sous formes d'imprimés et de graphiques.

ABSTRACT

A FORTRAN computer program, which implements an estimation procedure for a polynomial regression model with a change point, is described. The assumptions of the model are analogous to those for a single polynomial but include the assumption that the polynomial, and possibly the variance, have changed at a point in the data set. The order of the data points must also be known. The degree of the polynomial in each segment and the point of change are estimated by an iterative procedure which involves the use of the marginal likelihood function for the change-point parameter and a test using the change in residual sums of squares with reduction of degree. The assumptions of the model can be examined using the residuals and the likelihood function. Examples are given to demonstrate the utility of the program and to show the output from both the printer and the plotter.

INTRODUCTION

Regression analysis is often used to provide information about the relationship between variables. In most applications, one regression relationship is fitted to a set of data, but situations arise where it is known that the relationship has changed during the period of data collection. The complexity of the analysis depends upon the degree of knowledge about the form of the relationship and the point of change. If the point at which the regression relationship changes is known, then the analysis consists of fitting regular least squares relationships to the separate segments, possibly subject to the constraint that the regression functions agree at the point of change. If the point of change is unknown, then the analysis is more difficult since the change point must be estimated as well as the regression parameters. The present report addresses this latter situation and describes an algorithm based on a procedure given by Esterby and El-Shaarawi (1981a).

The algorithm has been written for regression on one dependent variable. The usual assumptions of least squares regression are made within each segment of the data fitted by one regression function, plus the assumption that the regression function has changed at an unknown point. The form of the regression function is restricted to polynomials, and an ordering of the observations, such as a natural ordering in time or space, is required. Let (x_i, y_i)

be independent pairs of observations, where the index i denotes the order of observation. Then the model consists of polynomials of degree p and q, with the change from a polynomial of degree p to a polynomial of degree q occurring after the observation with $i=n_1$. This is given by

$$y_i = \sum_{j=0}^{p} \theta_{1j} x_i^j + e_{1i}, \qquad i = 1,2,\ldots,n_1$$

and

$$y_i = \sum_{j=0}^{q} \theta_{2j} x_i^j + e_{2i}, \qquad i = n_1+1, n_1+2,\ldots,n,$$

where $e_{1i}(i=1,2,\ldots,n_1)$ and e_{2i} $(i=n_1+1, n_1+2,\ldots,n)$ are two independent sequences of random variables which are assumed to be $N(0,\sigma_1^2)$ and $N(0,\sigma_2^2)$, respectively. The parameters $\underline{\theta}_1 = (\theta_{10},\theta_{11},\ldots,\theta_{1p})$, $\underline{\theta}_2 = (\theta_{20},\theta_{21},\ldots,\theta_{2q})$, σ_1^2, σ_2^2 and n_1 are unknown. In Sections 2 and 3 this model is denoted by $M_{ü}(p,q|n_1)$. Given p and q, the change-point parameter, n_1, is estimated from the marginal likelihood function of n_1. If p and q are also unknown, an itertive procedure is used to estimate p and q. The procedure combines the marginal likelihood function of n_1 and a test analogous to that for reduction of degree in a single polynomial. A maximum value for the degree of each polynomial is assumed. Let (p^*,q^*) be the pair of maximum values, then the total degree $(p + q)$, is reduced by 1 at each step in the iteration, until no further reduction is possible or $(p,q) = (0,0)$. The case where $\sigma_1^2 = \sigma_2^2$ is also

considered and this model is denoted by $M_e(p,q|n_1)$, where the subscripts e and u are used to distinguish models obtained under the different variance assumptions. There are differences between the marginal likelihood functions and the procedures for estimating p and q in the cases where the variances are assumed equal and unequal. Details of the methods are given in Esterby and El-Shaarawi (1981a) and a more detailed account of likelihood functions in Esterby and El-Shaarawi (1981b).

The methods of residual analysis, which are available for checking the assumptions of the model when a single function is fitted, can be used here by combining the residuals from each segment. In addition, examination of the likelihood function permits a check on the assumption of one point of change.

The major objective of this procedure is to estimate the point at which the regression relationship changes when it can be assumed that one change has occurred. No constraint on the values of the regression functions at the change point is imposed. Other objectives may require different analyses. If it is required to have a smooth curve fitted to the data, fitting either a higher order polynomial or a cubic spline to the data would be appropriate in many cases.

A description of the algorithm is given in the next section and examples of the use of the program in the third section. The FORTRAN code has not been listed because the program is very long. A copy of the program and test cases can be obtained by writing to the

author. The program has been written in FORTRAN V, using some IMSL subroutines, and implemented on a CDC Cyber 171 with the NOS operating system. Plotting programs were written for a CALCOMP 1036 plotter. The purpose here is to describe the features of the program, without reiterating the development of the procedure, and illustrate its use by means of examples, so that the reader, in need of such a program, can evaluate its usefulness for a particular application.

ALGORITHM

Many features have been included in the program to enable the user to simply and efficiently analyze one or more pairs of variables in one computer job, provided that all the variables are on the same file. The options which must be specified by the user are listed in Table 1. The program can estimate the point of change, n_1, given the degrees of the polynomials or estimate the degrees of the polynomials and the point of change, given the maximum value which the degree can assume in each segment. For each pair of variables, the user can specify either the single pair of degrees, (p,q), or up to five values for the maximum degrees (p*,q*). For the single pair (p,q) or for each pair (p*,q*), the user needs to specify the assumption about the equality of variances, whether plots are to be produced and the two critical significance levels for the tests of hypotheses. Different plots are produced when the independent variable is not in ascending or descending order.

Figure 1 gives the main features of the program and shows where the user-specified options determine which parts of the program are used. To understand the details of the calculations which are performed for each step shown in the flow chart, it is necessary to read the paper by Esterby and El-Shaarawi, 1981a. Several comments will help clarify the logic. The iterative procedure for estimation of p and q, first assumes particular values for p and q and estimates n_1 as $\hat{n}_1$. Then holding n_1 fixed at $\hat{n}_1$, a test is performed to determine if the total degree (p+q) can be reduced by 1. If the iteration continues, that is, if the degree was found to be (p+q-1), then n_1 is again estimated assuming the new pair of degrees. The two sequences shown in Figure 1, for the cases of equal and unequal variances, are further explained in Table 2, where the differences, due to the two cases, are listed. The estimate of n_1, $\hat{n}_1$, is the value of n_1 which maximizes the marginal likelihood function and this is determined by calculating the marginal likelihood function for n_1 equal to the index values shown in Table 2.

Several of the steps shown in Figure 1 were not included in the procedure given in the 1981a paper. To test for equality of variances (see A in Figure 1), it is assumed that the variances are unequal and the point of change is estimated using the marginal likelihood function for unequal variances. A standard F test for equality of variances is then applied to test the hypothesis that the residual variances from segments 1 and 2 are equal. During the

estimation of p and q, if the degree in one segment reaches 0, then it is necessary to test for reduction in the other segment only. The test becomes identical to testing for reduction of degree when a single polynomial is fit to a set of data. Thus the critical significance level α is used directly, whereas in the case of two possible ways in which the degree can be reduced by 1, $\sqrt{\alpha}$ is used.

Implementation of the algorithm also required several additional steps. To permit matrix inversion, the independent variable is standardized. Further calculations are required to obtain the regression parameters in terms of the original independent variable from the parameters in terms of the standardized variable. The formulae are given in the appendix. In the case of equal variances, when neither p nor q is 0, the test for reduction of degree uses the sum of the residual sums of squares in the denominator of the test statistic. Thus the lower and upper limits for n_1 can be p+1 and n-(q+1), resulting in the degrees of freedom for the residual sum of squares in one segment being equal to 0. However, if this occurs when the degree is 0 in the other segment, the case in which the test is based on the segment with non-zero degree only, a test statistic with denominator degrees of freedom equal to 0 results. The correction to this situation which has been implemented, is to adjust the corresponding limit for n_1 so that there is always one observation more than the number of parameters to be estimated, and to maintain this adjustment for the rest of the iteration. For equal variances and the

number of parameters equal to the number of observations in one segment, the residual sum of squares is not calculated but assigned its theoretical value of 0. A zero residual sum of squares can also occur if all the values of the dependent variable are equal in one segment. Assigning to the residual sum of squares the theoretical value of 0 in the case of unequal variances is not satisfactory, since the logarithm of the residual sum of squares for each segment is needed for the marginal likelihood function. For this case, the residual sum of squares in the segment with constant dependent variable is assigned the arbitrary value of 0.0000001. This has the effect of treating constancy in the same manner for equal and unequal variances. The user can circumvent this arbitrary decision simply by specifying that the constant end values be omitted.

The results which are printed include a summary of the steps in the estimation of p and q, and for the final model, the relative marginal likelihood function of n_1, the regression parameters, $\underline{\hat{\theta}}_1$ and $\underline{\hat{\theta}}_2$, the predicted values of y, $\underline{\hat{y}}$, and the residuals $\underline{\hat{e}}=\underline{y}-\underline{\hat{y}}$. The residual sums of squares, residual mean squares, total corrected sum of squares and R^2, as defined in the appendix, are also printed. The residual analysis (Draper and Smith, 1981) is largely graphical, consisting of a Q-Q plot to check the assumption of normality, and plots of the standardized residuals versus $\underline{x}$ and $\underline{\hat{y}}$. A test for runs in the residuals is also performed.

EXAMPLES

Examples were used to assess the performance of the procedure in previous publications (Esterby and El-Shaarawi, 1980, 1981a,b). For the examples given here, the use of the computer program, including its model checking capabilities, is emphasized. Example 1 is discussed in detail and the program's printer and plot output are shown. In the other three examples, the plots are limited to those which illustrate particular features of the example.

Example 1

The rise of Ambrosia pollen towards the surface of a sediment core can be used as a time marker when the corresponding date of forest clearing is known. The assumptions of one change and of a polynomial functional form for the regression relationship between concentration and depth are tenable. The data set used here was given by Robbins, Edgington and Kemp (1978) and consists of Ambrosia concentrations (grains g^{-1} of dry sediment) and sediment depth (cm) for 23 samples of a sediment core from Lake Ontario. The mid-point of the sample interval was used as sample depth. Sample order was assigned from top to bottom, with the surface sample as sample one. The values of the options used to run the computer program are given in Table 3, and the resulting computer listing and plots in Figures 2 and 3.

The printer output has extensive titles and should be easy to understand (Figure 2). The conclusion from the test for equality of variances applies to all successive results. The relative likelihood function, residual sums of squares and mean squares, fitted model and residuals are printed for the final model only. The last part of the output is a summary of the final model. The factor listed there must be applied to all residual sums of squares and means squares in the previous results.

The final model was $M_4(1,1|15)$, that is, two straight lines, with different coefficients for samples 1 to 15 and samples 16 to 23, and unequal variances in the two segments. Since pollen analysts generally describe the changes from the bottom to the top of the core, the above model is one of an increase in concentration at the rate of 0.10 grains $g^{-1}cm^{-1}$ from the bottom sample to sample 16, and, at a higher rate, 0.94 grains $g^{-1}cm^{-1}$ for samples 15 to the surface. Higher variation was also exhibited in the upper 15 samples, with $\hat{\sigma}_1^2=41.29$ and $\hat{\sigma}_2^2=3.20$. The rates and variances are given as the second parameter based on untransformed x and the residual mean square for each segment in Figure 2. The point of change, sample 15, is the sample for which the relative likelihood function attains its maximum value of one. The strength of the evidence that sample 15 is the change point, relative to all other samples permitted by the model, is obtained by examining the relative likelihood function, $R(n_1)$. In this case all samples in the closed interval $[12,17]$ have a relative

likelihood function ≥ 0.10. By quoting, for example, the 0.10 relative likelihood interval, a measure of the uncertainty of the estimate is given.

For the independent variable in ascending or descending order, four plots are produced. Figure 3a contains the plot of the data with the fitted curves superimposed and the relative likelihood function. The left vertical and lower horizontal axes are for the data and fitted curves and the right vertical and upper horizontal axes for the relative likelihood function for the change-point parameter, n_1. Note that only integer values of n_1 can have non-zero relative likelihood, but the points for the relative likelihood function have been joined for ease of separation of the two types of information in the plot. The residual plots, shown in Figures 3b, c, and d, indicate no marked departures from the assumptions of the model. The standardized residual of largest magnitude is that of sample 6 at -2.5, and under normality one expects about 98% of the residuals will be in the interval $[-2.5, 2.5]$. Further, the observed significance level of 0.97 for the runs test (Figure 2) provides no evidence against the model assumption of independence.

The broad likelihood interval for the point of change would be of concern in using the Ambrosia horizon to date this core. As discussed by Robbins et al., a clear definition of what is meant by the Ambrosia rise or horizon has not been given. It is interesting to note, that if it is defined as the point at which the concentration

rises above background levels, then this means that q=0 in the present method. When (p,q) was specified as (1,0), the model which resulted (Figure 3e) was $M_u(1,0|17)$ with a 0.10 relative likelihood interval of $\lceil 15, 18 \rceil$. There were no marked differences in the residual analysis and $100R^2$=86.0 and 86.8 for $M_u(1,1|15)$ and $M_u(1,1|17)$, respectively. In Robbins et al., the Ambrosia rise had been set at sample 15 by inspection of the data. This corresponds to the first model estimated here, that of a change from one linear increase to another, with higher rate. Comparison of results from the present procedure with the poorly defined visual method should help the analyst establish a better defined procedure for determining Ambrosia horizons.

In general, the specification of (p*,q*) or the decision to assume (p,q) known depends upon the particular situation. In the Ambrosia example, to assume higher values for (p*,q*) would have resulted in fitting smaller excursions in the data and failed to address the question of a rise due to increased local input of Ambrosia pollen to the lake sediments. However, from the author's experience in analyzing other cores, it is clear that no one assumption will be appropriate for all cores.

Example 2

In the previous example, the order of the independent variable and the order of observation were the same and thus the two

regression models were based on non-overlapping ranges for the independent variable. Although this is the more commonly encountered case, situations arise where the ranges of the independent variable overlap. For example, such a situation could occur in the study of the stability of the relationship between two variables over time where the value of the independent variable is not controlled and at some point a different person started making the measurements.

The artificial data generated by Quandt (1958) was analyzed using the options specified in Table 3. The purpose here is to show the two plots which are produced for the data and fitted models and for the relative likelihood function (Figure 4). Different symbols are used for the data points and fitted lines in the two segments and the relative likelihood function is plotted in the same manner as in the case where only one plot is produced. The model corresponding to the plot is $M_e(1,1|12)$, with $\hat{y} = 2.22 + 0.69x$ and $\hat{y} = 5.91 + 0.48x$ being the fitted models for samples 1 to 12 and 13 to 20, respectively.

Example 3

The procedure does not include a test for the existence of a change. It assumes that one occurs. To show what might be expected when the procedure is applied to data in which no change has occurred,

the program was run on a sample of 100 random variables generated from a N(0,1) distribution. The program options are listed in Table 3 and the plots are shown in Figure 5. Since all possible values for the change-point parameter have a value of the relative likelihood function ≥ 0.30, clearly the correct conclusion, of no point of change, would be drawn.

Example 4

Although the method is designed for estimating only one point of change, several ad-hoc procedures are possible when more than one change occurs. One of the ad-hoc procedures is shown here using the percentages of Pinus pollen in a sediment core from Bog D Pond, Hubbard County, Minn. The percentages were calculated from the counts reported by McAndrews, 1963. The percentage of Pinus banksiana/resinosa-type declined sharply from high values in the lowest samples, whereas Pinus strobus increased towards the surface of the core. The sum of these two percentages should thus exhibit two changes. The program was run three times to reach a final model for the sum of the two percentages. The steps were as follows:

1. All 49 samples were used and the model $M_e(2,1|44)$ was obtained.
2. From samples 1 to 44, the model was $M_u(0,0|16)$.
3. To ascertain that the lower point of change is 44, a sequence containing only one change, namely samples 17 to 49 was analyzed. The result was the model $M_u(0,1|44)$.

The program options are given in Table 3. Since the same scales were specified in the plots for the last two runs, composite plots have been constructed by cutting and pasting together the regions 1 to 16 and 17 to 49 as shown in three of the plots (Figure 6 a, c and d). The separate Q-Q plots are given in Figure 6b. By picking the appropriate values from the last two runs, the final model is given by

$$\begin{aligned} \hat{y}_i &= 8.46 && i = 1,2,\ldots,16 \\ &= 1.25 && i = 17,18,\ldots,44 \\ &= 273 - 0.34\, x_i && i = 45,46,\ldots,49, \end{aligned}$$

with $\hat{\sigma}_1^2$=16.52, $\hat{\sigma}_2^2$=1.27 and $\hat{\sigma}_3^2$=2.62. Analogous to the calculation of $100R^2$ for two segments, the calculation for three segments leads to $100R^2$=75.5. Although some smaller peaks still appear in the residuals, the large changes have been located. Further, the observed significance level for the runs test based on the three-segment model is 0.20. The numbers required for calculating $100R^2$ and the significance level for the runs test were taken from one of the three program runs.

It is not always possible to obtain a satisfactory estimate of one of the points of change when more than one exists. This can occur if the changes are of similar magnitude. In this case, if segments of the data which contin only one change can be determined by examination of a scatter plot, a series of runs of the program will often provide satisfactory estimates.

DISCUSSION

The upper bounds for the degrees of the polynomials must be chosen carefully to prevent fitting small fluctuations in the data. As already emphasized, the fitted model should be examined for departures from the assumptions. The estimation of the point of change is sensitive to single points which are markedly different from surrounding points if the maximum value of the degree in one or both segments has been set too high. This will be obvious when the plots are examined.

The allowance for unequal variances in the two segments is an important feature in the procedure, since often when a system is disturbed the variability increases. Differences in the variability about the fitted curve in the two segments will not appear in the residual plots because standardized residuals are used. Changes in variability that do not take the simple form of constant variance within the segments can be detected from the residual plots and runs test.

The examples have not shown how the program is used for more than one pair of variables. If the three cases had been run separately in Example 4 without requesting plots, one run could then have been done for the three cases with plots requested. The same two columns of data would be specified for each case. A more unusual situation would be to specify different pairs of columns. For example

4, the columns might have corresponded to the three pairs: 1) depth and Pinus-banksiana/resinosa-type, 2) depth and Pinus strobus-type and 3) depth and the sum of Pinus-banksiana/resinosa-type and Pinus strobus-type.

APPENDIX

Standardization of the independent variable by subtraction of the mean and division by the standard deviation, where the mean and standard deviation are calculated from all n values, has proven sufficient. The standardization which is analogous to the use of the correlation matrix in the calculations for a single regression function is to standardize each segment separately. However, this would increase the calculations substantially since the samples in the segment change as the different possible values of n_1 are used. Thus the model that is fitted in one segment is

$$y_i = \beta_0 + \beta_1 \left(\frac{x_i - \bar{x}}{s} \right) + \ldots + \beta_d \left(\frac{x_i - \bar{x}}{s} \right)^d$$

and to obtain $\theta_0, \theta_1, \ldots, \theta_d$ which correspond to the model in terms of the original variables, $\underline{\theta} = (\theta_0, \theta_1, \ldots, \theta_d)'$ is calculated from

$$\underline{\theta} = A\underline{\beta}^* \quad \text{where } \underline{\beta}^* = \begin{bmatrix} \beta_0 \\ \beta_1/s \\ . \\ \beta_d/s^d \end{bmatrix}.$$

A is a (d+1) x (d+1) matrix which must be generated. The columns consist of terms of the binomial series $(1-\bar{x})^{j-1}$ for j = 1,2,...,d+1. The first j elements of the jth column of A are given by the terms of the series, in reverse order, and the last (d-j+1) elements are zero. For example, if d=2 then

$$\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 & -x & x^2 \\ 0 & 1 & -2x \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1/s \\ \beta_2/s^2 \end{bmatrix}$$

The multiple correlation coefficient, R^2, is defined as the regression sum of squares divided by the total sum of squares (TSS). In the present model, the numerator of R^2 is not just the sum of the regression sums of squares calculated separately for each segment, since this does not involve all of the components due to the model. To see this the total sum of squares can be decomposed as follows:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n_1} (y_i - \bar{y})^2 + \sum_{i=n_1+1}^{n} (y_i - \bar{y})^2$$

$$= TSS_1 + TSS_2.$$

But $$TSS_1 = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2 + \sum_{i=1}^{n_1} (\bar{y}_1 - \bar{y})^2$$

and the first term equals

$$\sum_{i=1}^{n_1} (y_i - \hat{y}_{1_i})^2 + \sum_{i=1}^{n_1} (\hat{y}_{1_i} - \bar{y}_1)^2,$$

where the subscript 1 refers to a quantity calculated from the samples in segment one. These two terms correspond to the residual and regression sums of squares for segment 1. By doing the same decomposition for segment two, recombining and grouping together the corresponding terms for both segments, it follows that

$$TSS = \{ResSS_1 + ResSS_2\} + \{RegSS_1 + RegSS^2\} + M$$

where $M = n_1(\bar{y}_1-\bar{y})^2 + (n-n_1)(\bar{y}_2-\bar{y})^2$ and ResSS and RegSS denote residual and regression sums of squares, respectively. Thus

$$100R^2 = 100 \{TSS - (ResSS_1 + ResSS_2)\}/TSS.$$

ACKNOWLEDGEMENTS

The algorithm was coded and the program tested by Mrs. J. Hodson. Plot subroutines were originally written by Mr. D. Pateman and an earlier version of a similar program by Mr. M. Fellowes.

REFERENCES

Draper, N.R. and H. Smith. 1981. Applied regression analysis, 2nd edn. John Wiley and Sons, Inc., New York.

Esterby, S.R. and A.H. El-Shaarawi. 1980. Examples of inferences about the point of change in a regression model. Unpublished report. National Water Research Institute, Burlington, Ontario, Canada.

Esterby, S.R. and A.H. El-Shaarawi. 1981a. Inference about the point of change in a regression model. Applied Statistics, 30(3):277-285.

Esterby, S.R. and A.H. El-Shaarawi. 1981b. Likelihood inference about the point of change in a regression regime. J. of Hydrology, 53:17-30.

McAndrews, J.H. 1966. Postglacial history of prairie, savanna and forest in Northwestern Minnesota. Mem. Torrey Bot. Club, 22(2), 72 p.

Quandt, R.E. 1958. The estimation of parameters of a linear regression system obeying two separate regimes. J. Amer. Statist. Ass., 53:873-880.

Robbins, J.A., D.N. Edgington and A.L.W. Kemp. 1978. Comparative ^{210}Pb, ^{137}Cs, and pollen geochronologies of sediments from Lakes Ontario and Erie. Quaternary Research, 10(2):256-278.

TABLE 1. User-specified program options

Number of pairs of variables

For each variable pair:

- Format for reading variable values, x then y
- Title
- Whether the independent variable is ordered
- Either 1) the known values of (p,q), or
  2) the number of pairs of maximum degree for the polynomials, n_{pq}, and their values (p*,q*)
- Whether all the samples are to be used or which ones are to be omitted
- Units for the axes of the plots

For the single pair (p,q) or for each pair of (p*,q*):

- Assumption about the variances (equal, unequal or test for equality)
- Whether plots are to be produced
- Critical significance level for the test for equality of variances
- Critical significance level for the test for reduction of degree

TABLE 2 Differences due to the assumption about the variances

	Equal Variances	Unequal Variances
Permissible values of n_1	$p+1,\ldots,n-(q+1)$	$p+2,\ldots,n-(q+2)$
Marginal likelihood function of n_1, $L(n_1)$	$\{\nu_1\hat{\sigma}_1^2+\nu_2\hat{\sigma}_2^2\}^{-1/2(\nu_1+\nu_2-1)}$	$\frac{\Gamma(1/2\,\nu_1)\ \Gamma(1/2\,\nu_2)}{\nu_1^{1/2\nu_1}\ \nu_2^{1/2\nu_2}}\ \hat{\sigma}_1^{-(\nu_1-1)}\ \hat{\sigma}_2^{-(\nu_2-1)}$
Test for reduction of degree, neither p=0 nor q=0	Residual sums of squares from the two segments are pooled to estimate the common variance.	Residual sum of squares for each segment is used separately to estimate the different variances.
Standardized residuals	Residuals from both segments are divided by $\hat{\sigma}$ obtained from the pooled residual sums of squares.	Residuals are standardized first by dividing by $\hat{\sigma}_1$ or $\hat{\sigma}_2$ and then combined.

$\nu_1 = n_1-(p+1)$, $\nu_2 = (n-n_1) - (q+1)$, $\hat{\sigma}_1^2 = \sum_{i=1}^{n_1} (y_i - \sum_{j=0}^{p} \hat{\theta}_{1j}\, x_i^j)^2/\nu_1$,

$\hat{\sigma}_2^2 = \sum_{i=n_1+1}^{n} (y_i - \sum_{j=0}^{q} \hat{\theta}_{2j}\, x_i^j)^2/\nu_2$ and $\Gamma(z)$ is the gamma function of z.

TABLE 3 Values of program options specified for the examples

	Example						
	1a	1b	2	3	4a	4b	4c
x ordered	yes	yes	no	yes	yes	yes	yes
(p,q) or (p*,q*)	(1,1)	(1,0)	(1,1)	(0,0)	(3,3)	(1,1)	(0,1)
Samples used	all	all	all	all	all	1-44	17-49
Variances	test	test	test	equal	test	test	test
α, variances	0.05	0.05	0.05	0.05	0.05	0.05	0.05
α, runs	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Plots	yes	yes	yes	yes	no	yes	yes
x origin	0.0	0.0	0.0	0.0		0.0	0.0
limit	68.4	68.4	18.0	100.0		800.0	800.0
increment	7.6	7.6	2.25	10.0		100.0	100.0
y origin	0.0	0.0	0.0	-3.3		0.0	0.0
limit	47.5	47.5	20.04	3.3		20.4	20.4
increment	9.5	9.5	3.34	1.1		3.4	3.4

a,b,c refers to runs within the example, thus 1a is the first program run described in Example 1.

FIGURE CAPTIONS

Figure 1 Flow chart showing the main features of the program

Figure 2 Printer output for first run of core KB, Example 1

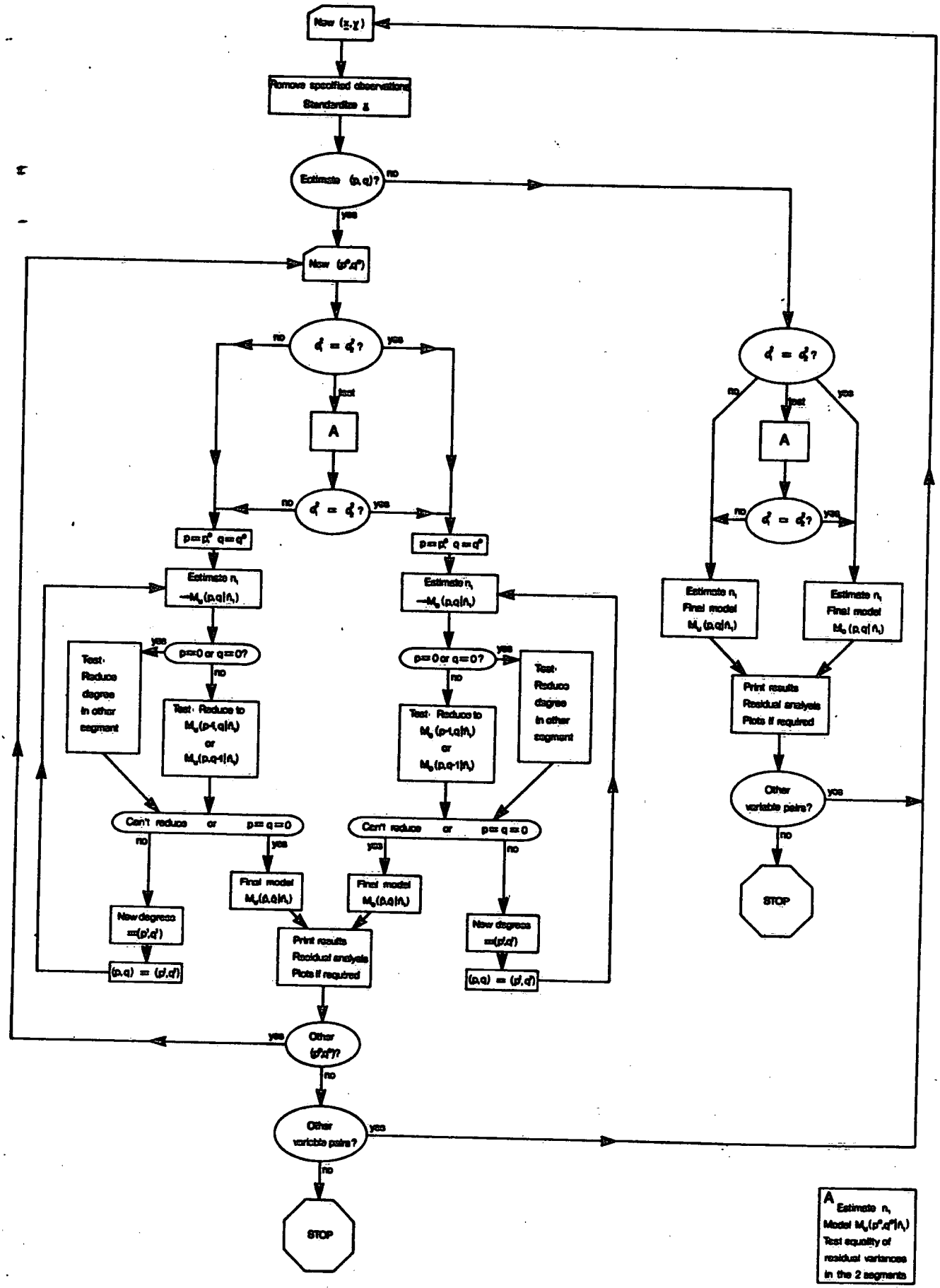
Figure 3 Plots for core KB, Example 1

Figure 4 Plot of the data with fitted curves and the relative likelihood function, Example 2

Figure 5 Plots for the normal data, Example 3

Figure 6 Plots constructed from runs 2 and 3 of Example 4

Fig 1



New (x,y)
Remove specified observations
Standardize x
Estimate (p,q)?
no
yes
New (p°,q°)
no
yes
test
A
no
yes
p=p° q=q°
Estimate n₁
p=0 or q=0?
yes
no
Test: Reduce degree in other segment
Test: Reduce to
or
Can't reduce or p=q=0
no
yes
Final model
New degrees =(p',q')
(p,q) = (p',q')
Print results
Residual analysis
Plots if required
Other (p°,q°)?
Other variable pairs?
STOP
Estimate n₁
Final model
Print results
Residual analysis
Plots if required
Other variable pairs?
STOP
A Estimate n₁
Model
Test equality of residual variances in the 2 segments

Fig 2

S.E.G.R.E.G

DATA SET 1

FORMAT FOR READING X AND Y: (TR6,F6.2,TL11,F5.1)
CORE KB AMBROSIA
X IS IN EITHER INCREASING OR DECREASING ORDER
P AND Q WILL BE ESTIMATED
23 PAIRS OF (X,Y) VALUES READ:

.500	46.600
1.500	31.900
3.000	37.500
5.000	30.000
7.000	34.400
9.000	12.900
11.000	38.300
13.000	24.300
15.000	20.400
17.000	17.400
19.000	27.000
21.000	24.000
23.000	12.900
25.000	14.400
27.000	11.400
29.000	6.000
35.000	6.400
42.500	2.100
47.500	2.600
52.500	1.300
57.500	2.000
62.500	4.900
67.500	1.200

DATA IS NOT MANIPULATED
MEAN OF X VALUES = 25.69565
STANDARD DEVIATION OF X VALUES = 20.36359
8 ADJACENT EQUAL VALUES AT BEGINNING OF Y VECTOR
8 ADJACENT EQUAL VALUES AT END OF Y VECTOR
PLOT SPECIFICATIONS
XORIGIN= 0.000 XLIMIT= 68.400 XSTEP= 7.600 YORIGIN= 0.000 YLIMIT= 47.500 YSTEP= 9.500

PAIR 1

MAXIMUM P = 1 MAXIMUM Q = 1
EQUALITY OF VARIANCES TO BE TESTED
PLOTS WILL BE PRODUCED
SIGNIFICANCE LEVEL FOR EQUALITY OF VARIANCES TEST = .05000
SIGNIFICANCE LEVEL FOR DETERMINATION OF DEGREES OF POLYNOMIALS = .01000

TEST FOR EQUALITY OF VARIANCES

SEGMENT	DEGREES OF FREEDOM	RESIDUAL SS	RESIDUAL MS
1	13	5.3674925157	.4128840397
2	6	.1920299883	.0320048480

F RATIO = 12.90067
SIGNIFICANCE LEVEL = .00246486
VARIANCES ARE UNEQUAL

ESTIMATION OF P AND Q, UNEQUAL VARIANCES

N1HAT	MODEL 1 P	MODEL 1 Q	MODEL 2 P	MODEL 2 Q	INCREASE IN RSS	F RATIO	OF	OBS S.L.	MAX S.L.	THEOR S.L.	CONCLUDE
15	1	1	0	1	9.496601	23.00065	1,13	.00035			
			1	0	.122858	3.83874	1, 6	.09779	.09779	.10000	ADEQUATE

THE THEORETICAL SIGNIGICANCE LEVEL IS THE SQUARE ROOT OF SPECIFIED SIGNIGICANCE LEVEL IF THERE ARE TWO POSSIBLE WAYS TO REDUCE THE DEGREE; OTHERWISE IT IS THE SPECIFIED SIGNIFICANCE LEVEL.

TABLE OF RESIDUAL SUMS OF SQUARES AND MEAN SQUARES CALCULATED DURING THE ESTIMATION OF P AND Q

N1HAT	P	Q	RSS1	RSS2	RMS1	RMS2
15	1	1	5.3674925157	.1920299883	.4128840397	.0320048480
15	0	0	14.8640933333	.3148875098	1.0617209524	.0449839286

Fig 2 cont'd

P = 1 Q = 1

N1	RESIDUAL SUMS OF SQUARES SEGMENT 1	SEGMENT 2	RESIDUAL MEAN SQUARES SEGMENT 1	SEGMENT 2	LIKELIHOOD FUNCTION (RELATIVE)
3	.8047605264	5.426092268	.8047605264	.3868960338	.0135801826
4	.8095717192	5.266771891	.5047605264	.3576618608	.0066346846
5	1.152913898	5.118371344	.3843713449	.3412247563	.0035910211
6	1.156475232	4.129010157	.2890688081	.2752673439	.0020677984
7	1.491756093	3.950841306	.2487926822	.2634560666	.0017201441
8	1.508111675	3.778848802	.2151139584	.2519232535	.0015506396
9	1.536031362	3.615718016	.1920039202	.2410478677	.0014359737
10	1.703110523	3.451214372	.1703110523	.2300809582	.0009508927
11	1.838242976	3.291278484	.1671129978	.2194185656	.0012150539
12	2.268924176	3.122210625	.1890770147	.2081473750	.0012680197
13	2.361278328	2.956	.1815983329	.1971	.0013764131
14	3.363444612	2.789	.2402460437	.1992	.0025528632
15	5.367492515	.192629	.3578328350	.0128	1.000000000
16	5.424266568	.1926	.3390166605	.0120	.3402528632
17	5.426183852	.1926	.3192931991	.0113	.3219195922
18	5.746114881	.192	.3192286045	.0106	.3074742161
19	5.839516934	.1926	.3073429312	.0101	.3052345418
20	5.773915209	.1926	.2886957605	.0096	.3014392642

N1HAT	RESS1	RESS2	RMS1	RMS2
15	5.367492515	.1926299883	.4129040397	.0128048488

I	V=F(THETA)	V=F(BETA)
1	36.94	36.94
2	1.50	1.50

ORDER OF OBSERVATION	X	OBSERVED Y	ESTIMATED Y	RESIDUAL	ORDERED STANDARDIZED RESIDUAL	STANDARD NORMAL QUANTILE
1	.5000000000	46.00000000	36.93223	9.06776	-2.49377	-2.01206
2	1.500000000	31.00000000	35.09577	-4.09577	-1.12676	-1.51881
3	2.500000000	37.00000000	34.25811	2.74188	-.754	-1.13364
4	3.500000000	30.00000000	33.42045	-3.42045	-.94078	-.84271
5	4.500000000	30.00000000	32.58279	-2.58279	-.71042	-.56373
6	5.500000000	34.00000000	31.74513	2.25486	.62028	-.33015
7	6.500000000	30.00000000	30.90747	-.90747	-.24961	-.11045
8	7.500000000	28.00000000	30.06981	-2.06981	-.56938	.11045
9	8.500000000	31.00000000	29.23215	1.76784	.48634	.33015
10	9.500000000	24.00000000	28.39449	-4.39449	-1.20887	.56373
11	10.50000000	28.00000000	27.55683	.44316	.12190	.84271
12	11.50000000	27.00000000	26.71917	.28082	.07724	1.13364
13	12.50000000	24.00000000	25.88151	-1.88151	-.51758	1.51881
14	13.50000000	25.00000000	25.04385	-.04385	-.01206	2.01206
15	14.50000000	23.00000000	24.20619	-1.20619	-.33178	
16	15.50000000	23.00000000	23.36853	-.36853	-.10138	
17	16.50000000	27.00000000	22.53087	4.46912	1.22896	
18	17.50000000	19.00000000	21.69321	-2.69321	-.74085	
19	18.50000000	27.00000000	20.85555	6.14444	1.69045	
20	19.50000000	17.00000000	20.01789	-3.01789	-.83012	
21	20.50000000	24.00000000	19.18023	4.81976	1.32565	
22	21.50000000	12.00000000	18.34257	-6.34257	-1.74501	
23	22.50000000	19.00000000	17.50491	1.49508	.41126	
24	23.50000000	14.00000000	16.66725	-2.66725	-.73372	
25	24.50000000	16.00000000	15.82959	.17040	.04687	
26	25.50000000	15.00000000	14.99193	.00806	.00221	
27	26.50000000	14.00000000	14.15427	-.15427	-.04244	
28	27.50000000	11.00000000	13.31661	-2.31661	-.63729	
29	28.50000000	6.000000000	12.47895	-6.47895	-1.78232	
30	29.50000000	6.000000000	11.64129	-5.64129	-1.55227	
31	30.50000000	6.000000000	10.80363	-4.80363	-1.32122	
32	31.50000000	2.000000000	9.965970	-7.96597	-2.19141	
33	32.50000000	2.000000000	9.128312	-7.12831	-1.96090	
34	33.50000000	2.000000000	8.290653	-6.29065	-1.73039	
35	34.50000000	2.000000000	7.452994	-5.45299	-1.49988	
36	35.50000000	1.000000000	6.615335	-5.61533	-1.54494	
37	36.50000000	1.000000000	5.777676	-4.77767	-1.31443	
38	37.50000000	1.000000000	4.940017	-3.94001	-1.08392	
39	38.50000000	1.000000000	4.102358	-3.10235	-.85341	
40	39.50000000	1.000000000	3.264699	-2.26469	-.62290	
41	40.50000000	1.000000000	2.427040	-1.42703	-.39239	
42	41.50000000	1.000000000	1.589381	-.589370	-.16188	
43	42.50000000	1.000000000	.751722	.248277	.06823	
44	43.50000000	1.000000000	-.086936	1.086936	.29874	
45	44.50000000	1.000000000	-.924595	1.924595	.52925	
46	45.50000000	1.000000000	-1.762254	2.762254	.75976	
47	46.50000000	1.000000000	-2.599913	3.599913	.99027	
48	47.50000000	1.000000000	-3.437572	4.437572	1.22078	
49	48.50000000	1.000000000	-4.275231	5.275231	1.45129	
50	49.50000000	1.000000000	-5.112890	6.112890	1.68180	
51	50.50000000	1.000000000	-5.950549	6.950549	1.91231	
52	51.50000000	1.000000000	-6.788208	7.788208	2.14282	
53	52.50000000	1.000000000	-7.625867	8.625867	2.37333	
54	53.50000000	1.000000000	-8.463526	9.463526	2.60384	
55	54.50000000	1.000000000	-9.301185	10.30118	2.83435	
56	55.50000000	1.000000000	-10.13884	11.13884	3.06486	
57	56.50000000	1.000000000	-10.97650	11.97650	3.29537	
58	57.50000000	1.000000000	-11.81416	12.81416	3.52588	
59	58.50000000	1.000000000	-12.65182	13.65182	3.75639	
60	59.50000000	1.000000000	-13.48948	14.48948	3.98690	
61	60.50000000	1.000000000	-14.32714	15.32714	4.21741	
62	61.50000000	1.000000000	-15.16480	16.16480	4.44792	
63	62.50000000	1.000000000	-16.00246	17.00246	4.67843	
64	63.50000000	1.000000000	-16.84012	17.84012	4.90894	
65	64.50000000	1.000000000	-17.67778	18.67778	5.13945	
66	65.50000000	1.000000000	-18.51544	19.51544	5.36996	
67	66.50000000	1.000000000	-19.35310	20.35310	5.60047	
68	67.50000000	1.000000000	-20.19076	21.19076	5.83098	
69	68.50000000	1.000000000	-21.02842	22.02842	6.06149	
70	69.50000000	1.000000000	-21.86608	22.86608	6.29200	
71	70.50000000	1.000000000	-22.70374	23.70374	6.52251	
72	71.50000000	1.000000000	-23.54140	24.54140	6.75302	
73	72.50000000	1.000000000	-24.37906	25.37906	6.98353	
74	73.50000000	1.000000000	-25.21672	26.21672	7.21404	
75	74.50000000	1.000000000	-26.05438	27.05438	7.44455	
76	75.50000000	1.000000000	-26.89204	27.89204	7.67506	
77	76.50000000	1.000000000	-27.72970	28.72970	7.90557	
78	77.50000000	1.000000000	-28.56736	29.56736	8.13608	
79	78.50000000	1.000000000	-29.40502	30.40502	8.36659	
80	79.50000000	1.000000000	-30.24268	31.24268	8.59710	
81	80.50000000	1.000000000	-31.08034	32.08034	8.82761	
82	81.50000000	1.000000000	-31.91800	32.91800	9.05812	
83	82.50000000	1.000000000	-32.75566	33.75566	9.28863	
84	83.50000000	1.000000000	-33.59332	34.59332	9.51914	
85	84.50000000	1.000000000	-34.43098	35.43098	9.74965	
86	85.50000000	1.000000000	-35.26864	36.26864	9.98016	
87	86.50000000	1.000000000	-36.10630	37.10630	10.21067	
88	87.50000000	1.000000000	-36.94396	37.94396	10.44118	
89	88.50000000	1.000000000	-37.78162	38.78162	10.67169	
90	89.50000000	1.000000000	-38.61928	39.61928	10.90220	
91	90.50000000	1.000000000	-39.45694	40.45694	11.13271	
92	91.50000000	1.000000000	-40.29460	41.29460	11.36322	
93	92.50000000	1.000000000	-41.13226	42.13226	11.59373	
94	93.50000000	1.000000000	-41.96992	42.96992	11.82424	
95	94.50000000	1.000000000	-42.80758	43.80758	12.05475	
96	95.50000000	1.000000000	-43.64524	44.64524	12.28526	
97	96.50000000	1.000000000	-44.48290	45.48290	12.51577	
98	97.50000000	1.000000000	-45.32056	46.32056	12.74628	
99	98.50000000	1.000000000	-46.15822	47.15822	12.97679	
100	99.50000000	1.000000000	-46.99588	47.99588	13.20730	

SUM OF RESIDUALS, SEGMENT 1 .0000000000

SUM OF RESIDUALS, SEGMENT 2 .0000000000

TEST FOR RUNS IN RESIDUALS

NUMBER OF RUNS	16
NUMBER OF POSITIVE SIGNS	10
NUMBER OF NEGATIVE SIGNS	13
MEAN	12.3
STANDARD DEVIATION	2.38
N(0,1) VARIATE	1.62
PROBABILITY LEVEL	.97

FINAL MODEL

ALL RESIDUAL SUMS OF SQUARES AND RESIDUAL MEAN SQUARES WERE DIVIDED BY 100.

POINT OF CHANGE PARAMETER = 15

DEGREE OF POLYNOMIAL, SEGMENT 1 = 1

DEGREE OF POLYNOMIAL, SEGMENT 2 = 1

REGRESSION PARAMETERS

A. TRANSFORMED X

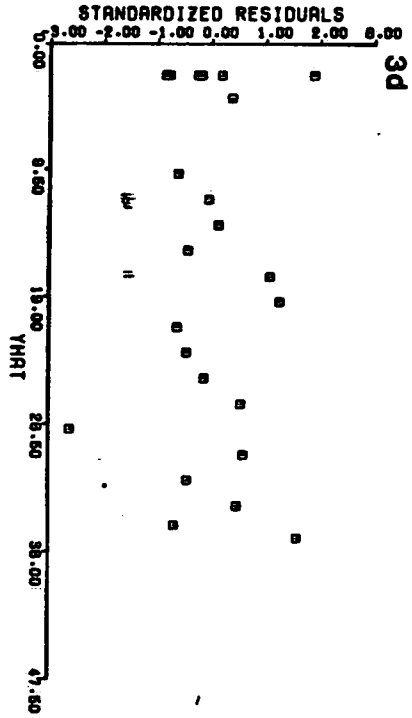
SEGMENT 1	13.18205	-19.28819	0.00000	0.00000	0.00000	0.00000
SEGMENT 2	5.64865	-2.81965	0.00000	0.00000	0.00000	0.00000

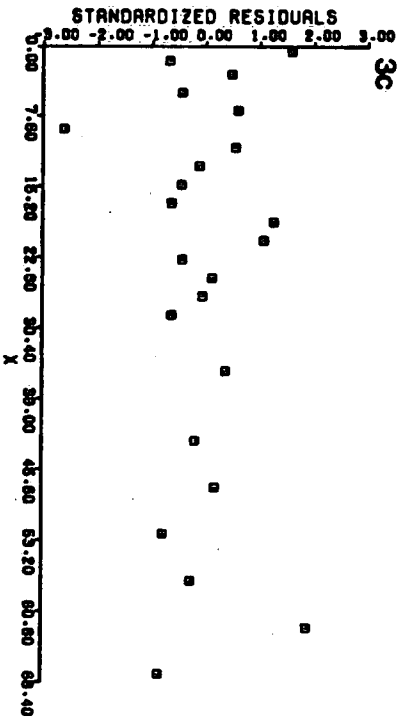
B. UNTRANSFORMED X

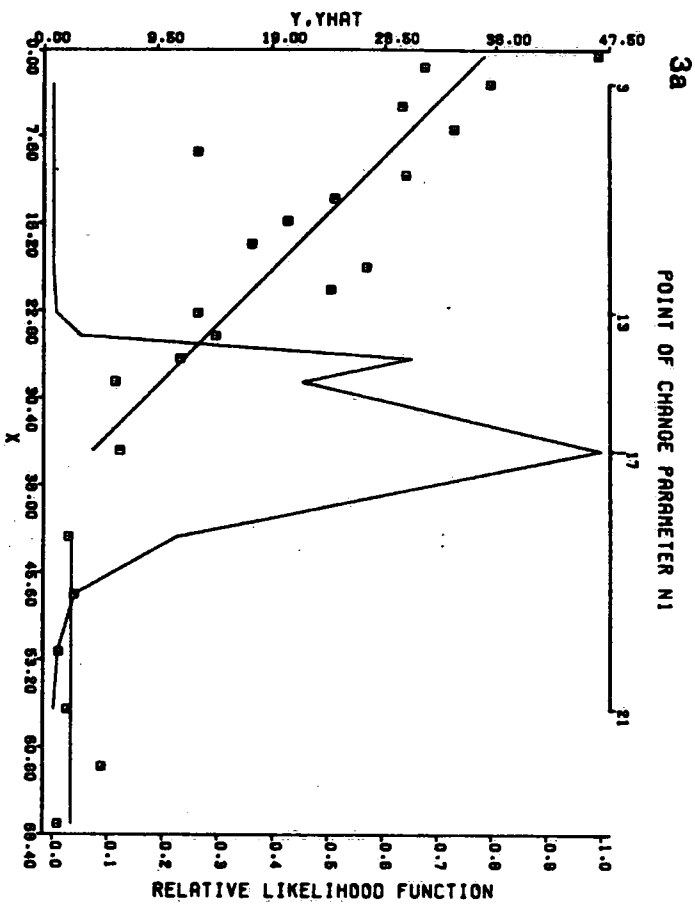
SEGMENT 1	37.48967	-.94287	0.00000	0.00000	0.00000	0.00000
SEGMENT 2	8.19788	-.09918	0.00000	0.00000	0.00000	0.00000

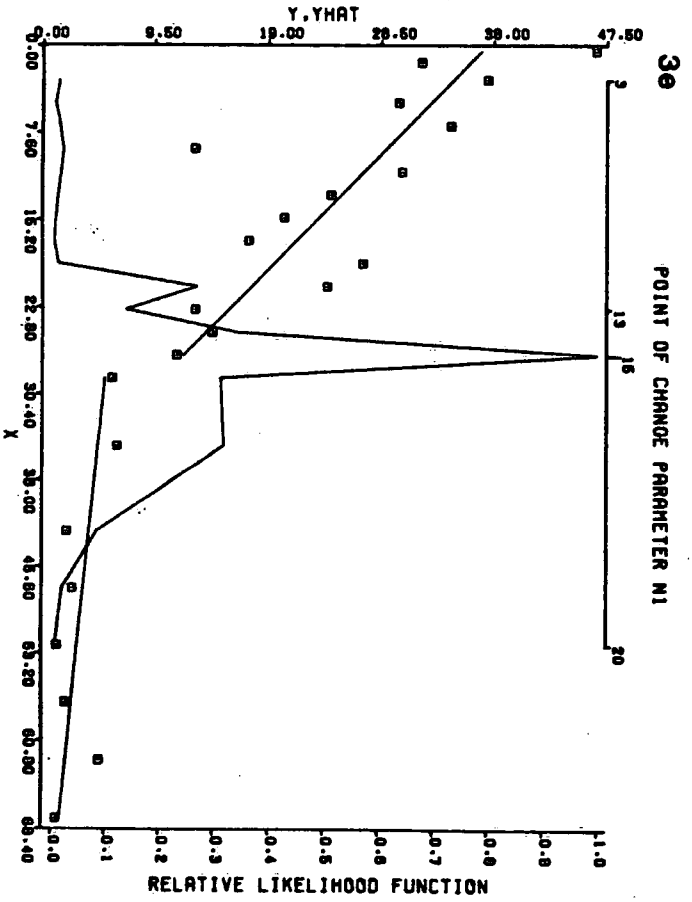
100 R SQUARED = 86.82406

TOTAL SUM OF SQUARES = 3977.92435


STANDARDIZED RESIDUALS
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
YHAT
0.00 9.50 19.00 28.50 38.00 47.50
3d


STANDARDIZED RESIDUALS
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
X
0.00 7.60 15.20 22.80 30.40 38.00 45.60 53.20 60.80 68.40
3c


Y,YHAT
0.00 9.50 19.00 28.50 38.00 47.50
X
0.00 7.60 15.20 22.80 30.40 38.00 45.60 53.20 60.80 68.40
POINT OF CHANGE PARAMETER N1
5 13 17 21
RELATIVE LIKELIHOOD FUNCTION
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
3a


Y,YHAT
0.00 9.50 19.00 28.50 38.00 47.50
X
0.00 7.60 15.20 22.80 30.40 38.00 45.60 53.20 60.80 68.40
POINT OF CHANGE PARAMETER N1
5 13 15 20
RELATIVE LIKELIHOOD FUNCTION
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
3e

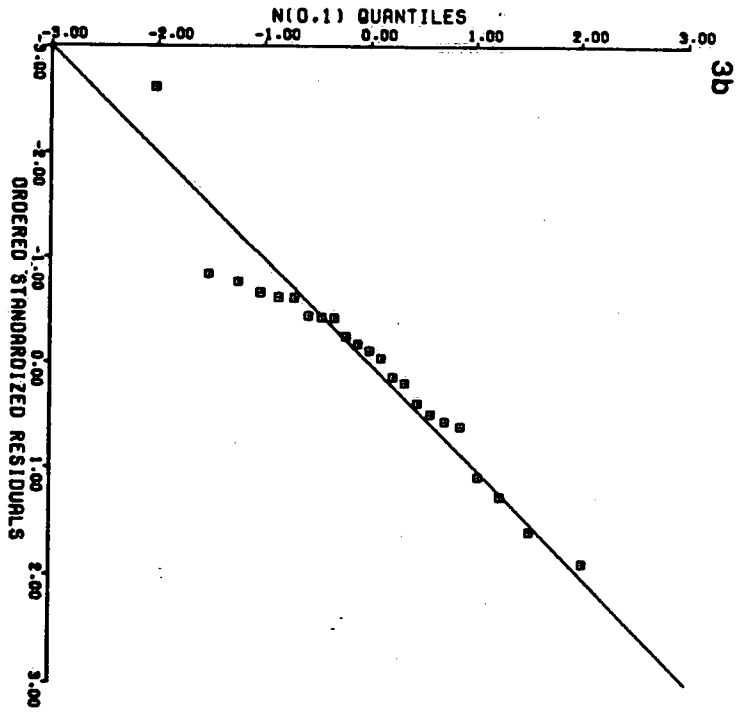
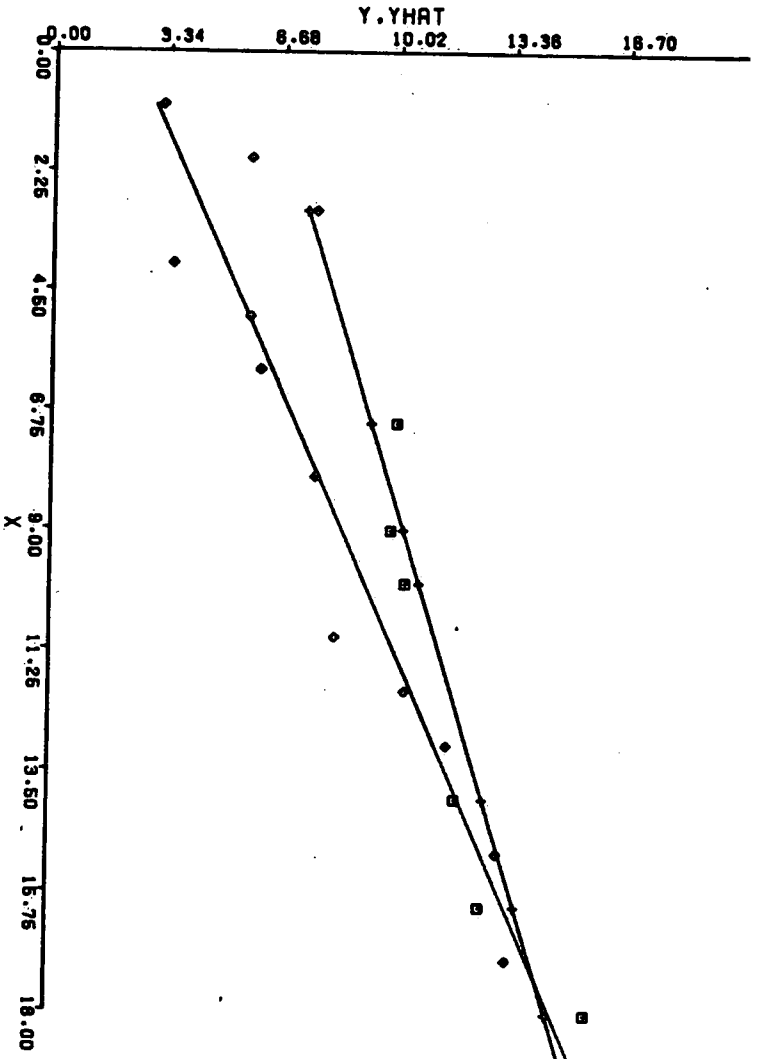
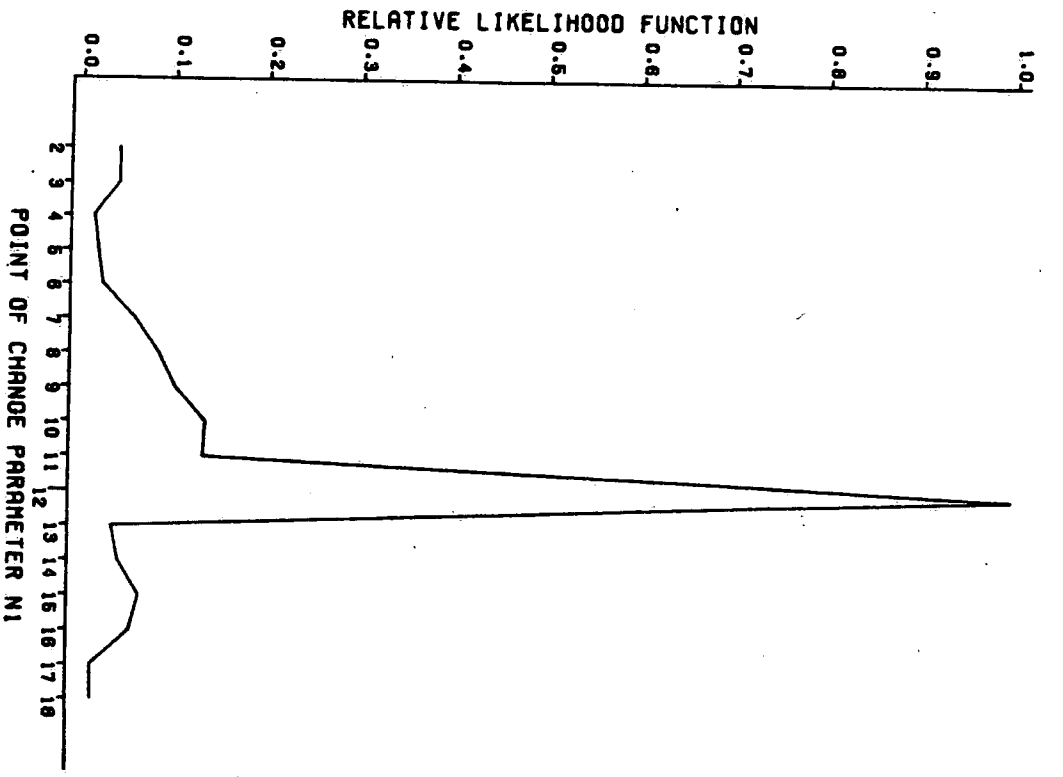
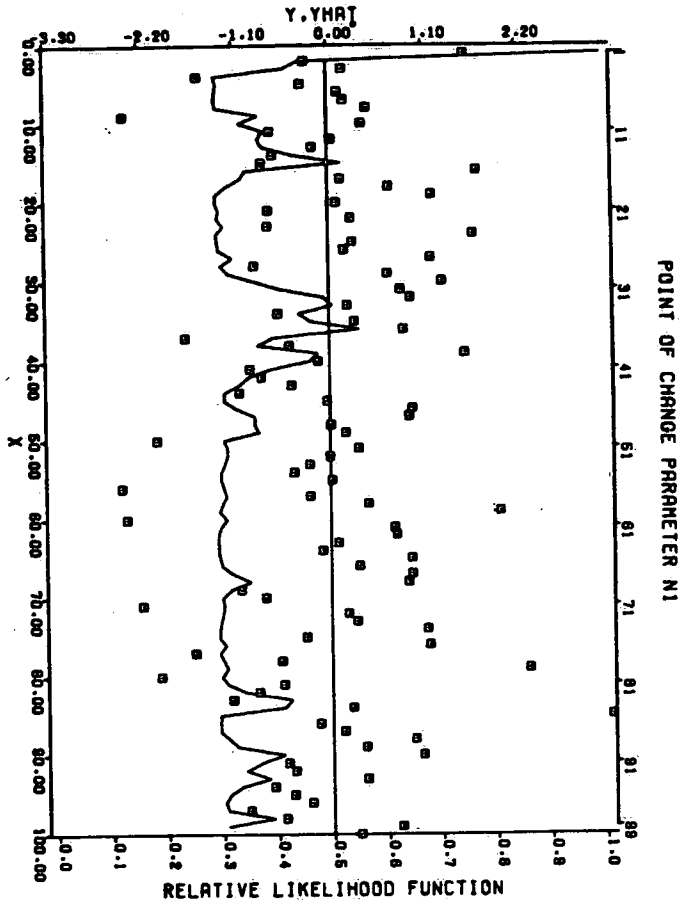

N(0,1) QUANTILES
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
ORDERED STANDARDIZED RESIDUALS
-3.00 -2.00 -1.00 0.00 1.00 2.00 3.00
3b

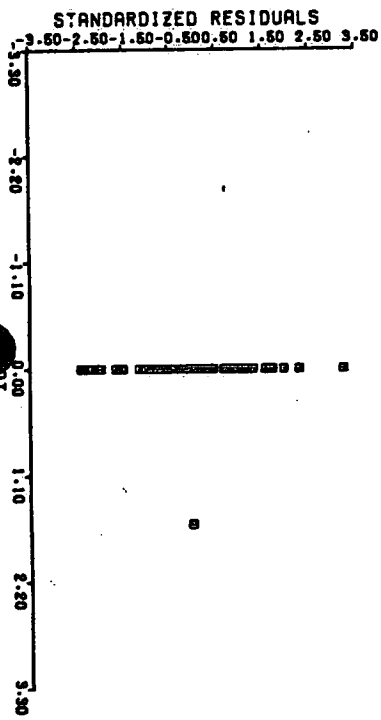
Fig 4

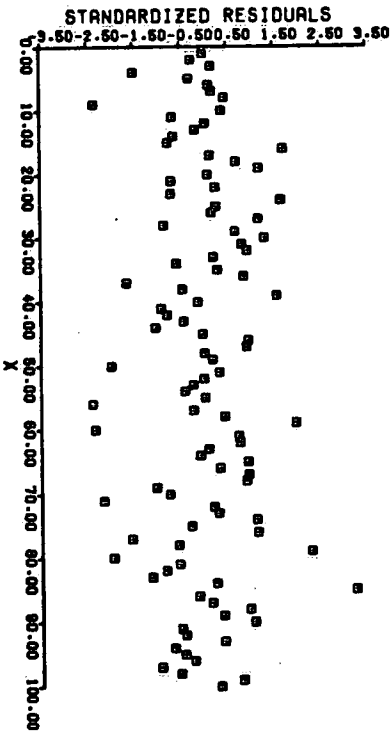

Y.YHAT
0.00
3.34
6.68
10.02
13.36
16.70
X
0.00
2.25
4.50
6.75
9.00
11.25
13.50
15.75
18.00

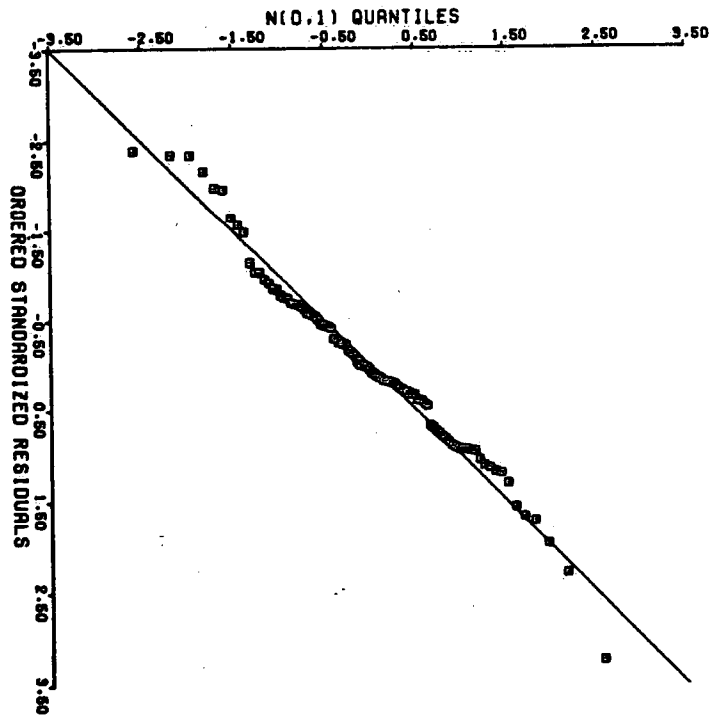

RELATIVE LIKELIHOOD FUNCTION
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
POINT OF CHANGE PARAMETER N1
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Figs

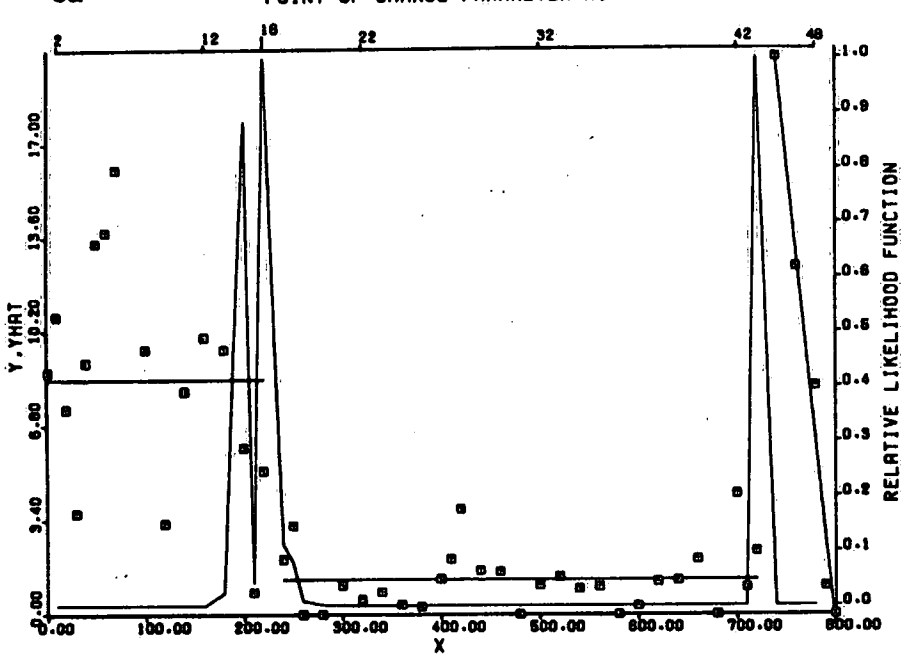

Y,YHAT
-3.30
-2.20
-1.10
0.00
1.10
2.20
X
POINT OF CHANGE PARAMETER N1
RELATIVE LIKELIHOOD FUNCTION


STANDARDIZED RESIDUALS
-3.50
-2.50
-1.50
-0.500
0.50
1.50
2.50
3.50

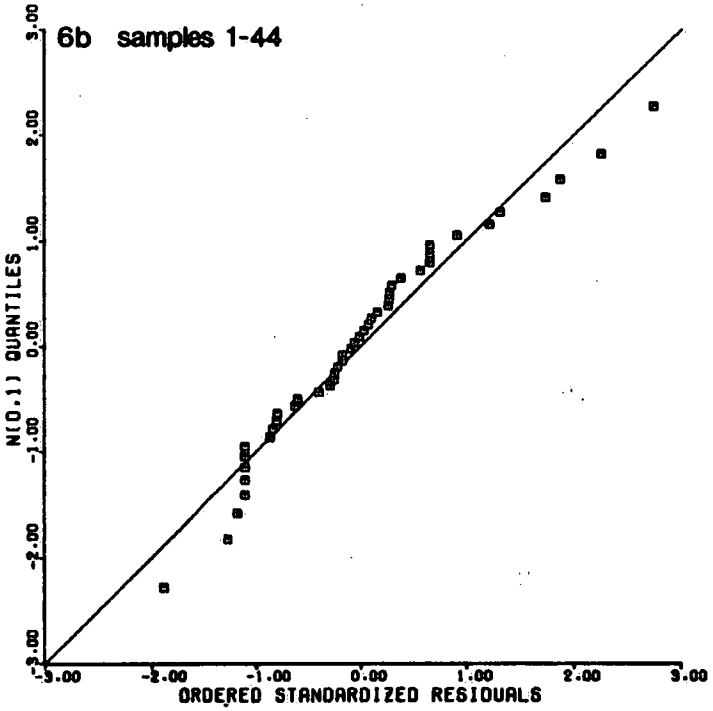

STANDARDIZED RESIDUALS
X


N(0,1) QUANTILES
ORDERED STANDARDIZED RESIDUALS

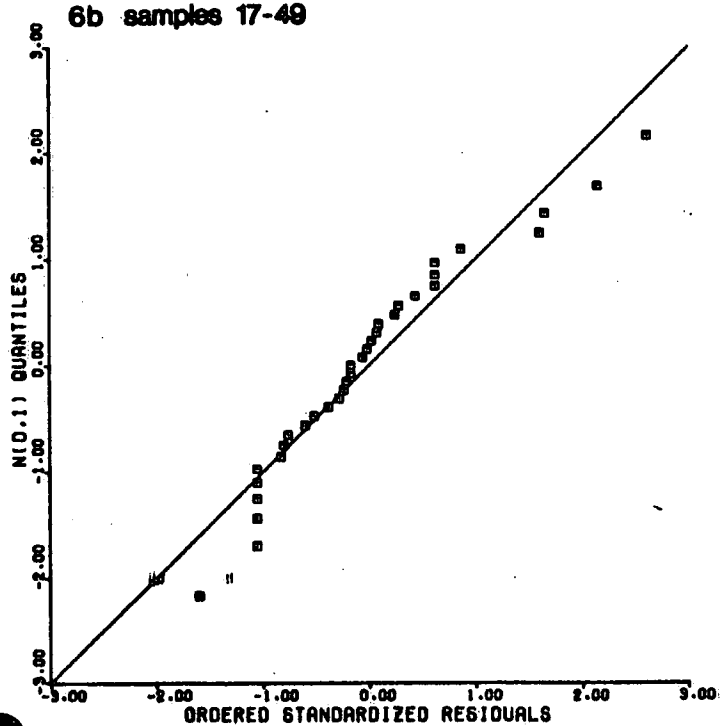
6a



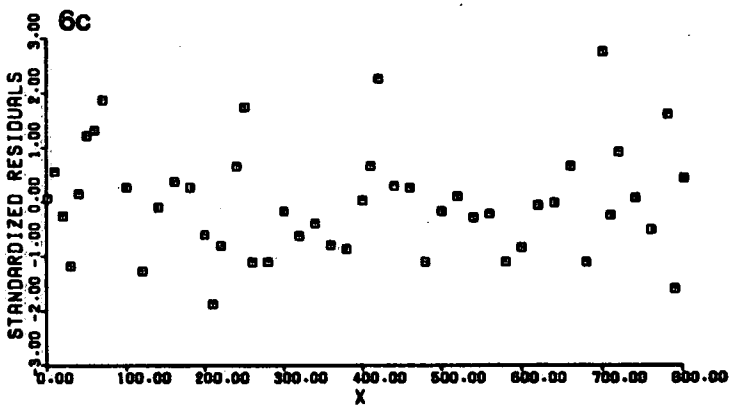
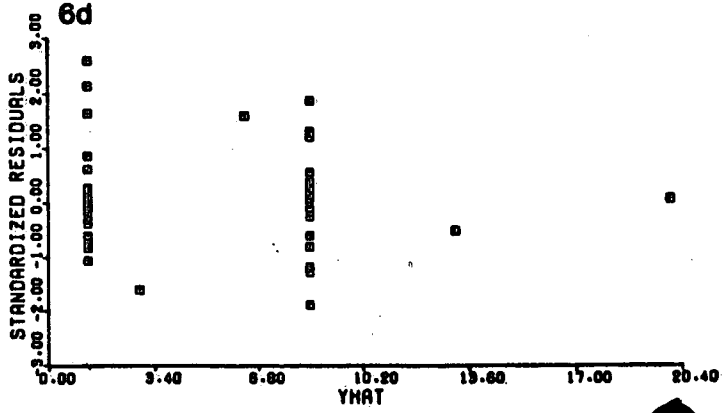
6b samples 1-44



6b samples 17-49



6c



6d