SPLIT-LINE REGRESSION TECHNIQUES

by

JOHN GLOWIK

B.Sc. (Hons.) - Adelaide

Thesis submitted for the degree of Master of Science at The University of Adelaide, Department of Mathematical Statistics, March, 1977.

CONTENTS

SUMMARY

The results of this thesis deal with a numerical investigation of data exhibiting split-line regression characteristics.

Chapter 1 gives a general introduction and a more informative description of split-line regressions.

In Chapter 2 we develop some procedures to test for significant non-linearity characteristics in data sets. In particular four methods are considered in detail and applied to examples in Chapter 5.

In Chapter 3 we consider some previously developed procedures for finding the least squares estimate for the intersection point of two regression lines. We also develop new methods of approach in order to find an estimate for the intersection point, γ.

In Chapter 4 we consider some inferential problems, where the asymptotic distribution and confidence interval for γ are discussed.

In Chapter 5 we look at some experimental and generated data sets, comparing with respect to time, the various methods considered for finding an estimate of γ.

In the Appendix we list, in FORTRAN IV, the routines which were developed and used for the various methods of approach.

SIGNED STATEMENT

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University. To the best of my knowledge and belief, the thesis contains no material previously published or written by any other person, except where due reference is made in the text of the thesis.

_____

(J

ACKNOWLEDGEMENTS

Dr. G.M. Tallis suggested the topic of this thesis to the author and offered valuable assistance. My sincere thanks to Dr. W.N. Venables and Mr. K.W. Morris for the many discussions and helpful comments. My sincere thanks are also due to Katherine Halsey for her careful and proficient mathematical typing.

Many thanks to my friends and colleagues in the Mathematical Statistics Department of The University of Adelaide for their encouragement and support throughout.

CHAPTER 1

In fitting linear models to data, it may be expedient to consider split line models. This can arise because of a-priori knowledge on the part of the experimenter, or from information obtained from a data plot.

Situations where split line models could occur with some frequency is in data representing the regression of growth measurement against time. For example situations where this could occur are,

(1) the onset of a disease resulting in a reduced growth rate,

(2) the application of a treatment having an immediate stimulating or inhibiting effect,

(3) physical injury of an organism,

(4) the occurrence of a rapid change in external conditions.

The research presented in this thesis is intended to bring together and compare various approaches for finding an estimate for the intersection point (γ) of two regression lines. Therefore we have given some attention to numerical comparisons of the various methods as well as finding efficient computational procedures for each approach.

The split-line regression model can be written as

$$E(Y_{ij}) = \begin{cases} \alpha_1 + \beta_1 x_i, & x_i \leq \gamma \\ \alpha_2 + \beta_2 x_i, & x_i \geq \gamma \end{cases}$$

where $i = 1,\ldots,M$; $j = 1,\ldots,m_i$ and the intersection point

$$\gamma = (\alpha_1 - \alpha_2)/(\beta_2 - \beta_1) \ .$$

CHAPTER 2

TESTING FOR A SPLIT-LINE REGRESSION MODEL

2.1 Introduction

In many cases there may be good initial reasons to expect a split-line regression model to apply. An examination of a data plot would usually indicate the existence of a split-line regression, and to demonstrate this we construct a test of the null hypothesis that there is just one regression line. Such tests have been developed by Page (11), Quandt (12), Bhattacharyya and Johnson (2), Brown and Durbin (3); and others.

2.2 Basic Results

Regression coefficients found by progressively larger subsets of the data are used in two tests developed. This recursive method of calculating updated regression coefficients makes use of the following.

R2.2.1 [Rao (13), page 33 point 2.8]

Let A be a non-singular matrix, and $\underset{\sim}{u}$, $\underset{\sim}{V}$ be two column vectors. Then

$$(A+\underset{\sim}{U}\underset{\sim}{V}')^{-1} = A^{-1} - \frac{A^{-1}\underset{\sim}{u}\underset{\sim}{V}'A^{-1}}{1+\underset{\sim}{V}'A^{-1}\underset{\sim}{u}}, \quad \underset{\sim}{V}'A^{-1}\underset{\sim}{u} \neq -1.$$

This gives a method of calculating $(A+\underset{\sim}{u}\underset{\sim}{V}')^{-1}$ from A^{-1}.

2.2.1 Updating Regression Coefficients

Allowing for multiple y readings on each distinct value in X, $X = [\underset{\sim}{x}; x_1 < x_2 < \ldots < x_M]$, let $m_j \geq 1$ be the multiplicity of observations at the value $x = x_j$. Suppose we have the least squares estimate, $\underset{\sim}{b}_k$, of $\underset{\sim}{\beta}' = (\alpha,\beta)$ for

the regression model $\underset{\sim}{y}_k = X_k\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_k$, based on the first k $(k>1)$ values of X. An additional m_{k+1} observations are made at $x = x_{k+1}$ and the least squares estimate of $\underset{\sim}{\beta}$ using all $k+1$ data points is required. The extended model can now be rewritten as,

$$\begin{pmatrix} \underset{\sim}{y}_k \\ \underset{\sim}{y}^*_{k+1} \end{pmatrix} = \begin{pmatrix} X_k \\ X^*_{k+1} \end{pmatrix} \underset{\sim}{\beta} + \begin{pmatrix} \underset{\sim}{\varepsilon}_k \\ \underset{\sim}{\varepsilon}^*_{k+1} \end{pmatrix} \quad \text{where,}$$

$\underset{\sim}{y}^*_{k+1}$: new set of m_{k+1} observations.

X^*_{k+1} : $\begin{bmatrix} 1 & x_{k+1} \\ \vdots & \vdots \\ 1 & x_{k+1} \end{bmatrix}$, a $m_{k+1} \times 2$ matrix.

$\underset{\sim}{\varepsilon}^*_{k+1}$: error sub-vector corresponding to $\underset{\sim}{y}^*_{k+1}$

$\underset{\sim}{y}_k$: vector of observations for the first k values of X.

Let $\underset{\sim}{b}_{k+1}$ denote the least-squares estimate of $\underset{\sim}{\beta}$ based on data from the first $k+1$ values of X. It follows that

$$\underset{\sim}{b}_{k+1} = [X_k'X_k + X^{*\prime}_{k+1}X^*_{k+1}]^{-1}[X_k'\underset{\sim}{y}_k + X^{*\prime}_{k+1}\underset{\sim}{y}_{k+1}].$$

By R2.2.1 it readily follows that

$$[X_k'X_k + X^{*\prime}_{k+1}X^*_{k+1}]^{-1}$$

$$= (X_k'X_k)^{-1} - \frac{m_{k+1}(X_k'X_k)^{-1}\underset{\sim}{x}_{k+1}\underset{\sim}{x}'_{k+1}(X_k'X_k)^{-1}}{1+m_{k+1}\underset{\sim}{x}'_{k+1}(X_k'X_k)^{-1}\underset{\sim}{x}_{k+1}} \quad , \text{ where } \underset{\sim}{x}'_{k+1} = (1, x_{k+1});$$

When $m_{k+1} = 1$, this result simplifies to that given by Kalman (9).

whence

$$\underset{\sim}{b}_{k+1} = (X_k'X_k)^{-1}X_k'\underset{\sim}{y}_k + (X_k'X_k)^{-1}\underset{\sim}{x}_{k+1}\underset{\sim}{1}'_{k+1}\underset{\sim}{y}^*_{k+1}$$

$$- m_{k+1}\frac{(X_k'X_k)^{-1}\underset{\sim}{x}_{k+1}\underset{\sim}{x}'_{k+1}\ (X_k'X_k)^{-1}[X_k'\underset{\sim}{y}_k + \underset{\sim}{x}_{k+1}\underset{\sim}{1}'_{k+1}\underset{\sim}{y}^*_{k+1}]}{1+m_{k+1}\underset{\sim}{x}'_{k+1}(X_k'X_k)^{-1}x_{k+1}}$$

$$= \underset{\sim}{b}_k + \frac{(X_k'X_k)^{-1}\underset{\sim}{x}_{k+1}[\underset{\sim}{1}'_{k+1}\underset{\sim}{y}^*_{k+1} - m_{k+1}\underset{\sim}{x}'_{k+1}\underset{\sim}{b}_k]}{1+m_{k+1}\underset{\sim}{x}'_{k+1}\ (X_k'X_k)^{-1}\underset{\sim}{x}_{k+1}}$$

With slight modification the above result also applies to multiple linear regression.

Putting

$$T_{k+1} = 1 + m_{k+1}\underset{\sim}{x}'_{k+1}(X_k'X_k)^{-1}\underset{\sim}{x}_{k+1}$$

$$\underset{\sim}{v}_{k+1} = (X_k'X_k)^{-1}\underset{\sim}{x}_{k+1} \quad , \quad \text{then}$$

$$\underset{\sim}{b}_{k+1} = \underset{\sim}{b}_k + \underset{\sim}{v}_{k+1}(\underset{\sim}{1}'_{k+1}\underset{\sim}{y}^*_{k+1} - m_{k+1}\underset{\sim}{x}'_{k+1}\underset{\sim}{b}_k)/T_{k+1}; \quad k = 2,\ldots,M$$

The main use of this result is to establish some statistical properties, since for simply linear regression it is simpler to invert the successive 2×2 matrix, (X_kX_k') in calculating $\underset{\sim}{b}_k$. These properties we now turn to.

R2.2.2

$$\text{Cov}\ (\underset{\sim}{b}_{k+S}, \underset{\sim}{b}_k) = \text{Var}\ (\underset{\sim}{b}_{k+S})$$

for $k = 2,\ldots M$; $S = 1,\ldots M-k$.

Let $\underset{\sim}{b}_k' = (\hat{\alpha}_k, \hat{\beta}_k)$, $\underset{\sim}{y}_k' = (\underset{\sim}{y}_1^{*\prime}\ldots\underset{\sim}{y}_k^{*\prime})$

and $$\ell = \sum_{i=k+1}^{k+S} m_i .$$

Using the result that

$$\text{Cov}\ (A\underset{\sim}{y}, B\underset{\sim}{y}) = A\ \text{var}\ (\underset{\sim}{y})B' \quad \text{and}$$

$$\underset{\sim}{b}_k = (X_k'X_k)^{-1}X_k'\underset{\sim}{y}_k$$

$$= [(X_k'X_k)^{-1}X_k', 0]\underset{\sim}{y}_{k+S} \ , \ 0 \text{ being a } 2 \times \ell \text{ matrix of zeros.}$$

then

$$\text{Cov}\ (\underset{\sim}{b}_{k+S}, \underset{\sim}{b}_k) = (X'_{k+S}X_{k+S})^{-1}X'_{k+S}\sigma^2 I_{k+S}\begin{pmatrix} X_k \\ 0 \end{pmatrix}(X_k'X_k)^{-1}$$

$$= \sigma^2(X'_{k+S}X_{k+S})^{-1}(X_k'X_k)(X_k'X_k)^{-1}$$

$$= \underline{\text{Var}\ (\underset{\sim}{b}_{k+S})}$$

R2.2.3

Without any loss of generality let $k > t$, then by use of R2.2.2 we show

$$\text{Cov}\ (\underset{\sim}{b}_{k+1} - \underset{\sim}{b}_k, \underset{\sim}{b}_{t+1} - \underset{\sim}{b}_t) = 0_{(2\times 2)}.$$

$$\text{L.H.S.} = \text{Cov}\ (\underset{\sim}{b}_{k+1}, \underset{\sim}{b}_{t+1}) - \text{Cov}\ (\underset{\sim}{b}_{k+1}, \underset{\sim}{b}_t) - \text{Cov}\ (\underset{\sim}{b}_k, \underset{\sim}{b}_{t+1}) + \text{Cov}\ (\underset{\sim}{b}_k, \underset{\sim}{b}_t)$$

$$= \text{Var}\ (\underset{\sim}{b}_{k+1}) - \text{Var}\ (\underset{\sim}{b}_{k+1}) - \text{Var}\ (\underset{\sim}{b}_k) + \text{Var}\ (\underset{\sim}{b}_k)$$

$$= \underline{0_{(2\times 2)}}$$

R2.2.4

$$\text{Var}\ (\underset{\sim}{b}_{k+1}-\underset{\sim}{b}_k) = \sigma^2\ \frac{m_{k+1}\underset{\sim}{v}_{k+1}\underset{\sim}{v}'_{k+1}}{T_{k+1}}$$

L.H.S. $= \text{Var}\ (\underset{\sim}{b}_{k+1}) + \text{Var}\ (b_k) - 2\ \text{Cov}\ (\underset{\sim}{b}_{k+1}, b_k)$

$= \text{Var}\ (\underset{\sim}{b}_k) - \text{Var}\ (\underset{\sim}{b}_{k+1})$, using R2.2.2

$= \sigma^2[(X'_kX_k)^{-1} - (X'_{k+1}X_{k+1})^{-1}]$

$= \sigma^2\ m_{k+1}\underset{\sim}{v}_{k+1}\underset{\sim}{v}'_{k+1}/T_{k+1}$, using 2.2.1

R2.2.5

$$\text{Rank}\ (\text{Var}[b_{k+1}-b_k]) = 1.$$

From R2.2.4 it follows that

$$\text{rank}\ (\text{Var}[\underset{\sim}{b}_{k+1}-\underset{\sim}{b}_k]) = \text{rank}\ (\underset{\sim}{v}_{k+1}\underset{\sim}{v}'_{k+1})$$

$$= 1$$

This means that the Variance matrix of $\underset{\sim}{b}_{k+1} - \underset{\sim}{b}_k$ is singular. Writing $(\underset{\sim}{b}_{k+1}-\underset{\sim}{b}_k)' = (\hat{\alpha}_{k+1}-\hat{\alpha}_k, \hat{\beta}_{k+1}-\hat{\beta}_k)$, it follows that $\hat{\alpha}_{k+1} - \hat{\alpha}_k$ and $\hat{\beta}_{k+1} - \hat{\beta}_k$ are linearly related, so that inferences based on $\underset{\sim}{b}_{k+1} - \underset{\sim}{b}_k$ can be expressed in terms of either difference. Let

$$\underset{\sim}{A}_{k+1} = \begin{pmatrix} a_{1,k+1} \\ a_{2,k+1} \end{pmatrix} = \left(\frac{m_{k+1}}{T_{k+1}}\right)^{\frac{1}{2}} \underset{\sim}{v}_{k+1}\ ,$$

then $\text{Var}(\underset{\sim}{b}_{k+1}-\underset{\sim}{b}_k) = \sigma^2\ \underset{\sim}{A}_{k+1}\underset{\sim}{A}_{k+1}'$ and to derive the linear relation, for $\text{Var}(\underset{\sim}{\lambda}'[\underset{\sim}{b}_{k+1}-\underset{\sim}{b}_k]) = 0$ this implies that $\lambda_1(\hat{\alpha}_{k+1}-\hat{\alpha}_k) + \lambda_2(\hat{\beta}_{k+1}-\hat{\beta}_k) = \text{Constant}$. For $\lambda_1 = a_{2,k+1}$, $\lambda_2 = -a_{1,k+1}$ we have $a_{2,k+1}(\hat{\alpha}_{k+1}-\hat{\alpha}_k) - a_{2,k+1}(\hat{\beta}_{k+1}-\hat{\beta}_k) = C$, and by taking expectations it follows that $C = 0$. Therefore

$$(\hat{\alpha}_{k+1}-\hat{\alpha}_k) = \left(\frac{a_{1,k+1}}{a_{2,k+1}}\right)\ (\hat{\beta}_{k+1}-\hat{\beta}_k);\quad a_{2,k+1} \neq 0$$

2.3 Tests using differences between consecutive updated regression coefficents.

We assume the errors ε_{ij} $(i=1,\ldots,M;\ j=1,\ldots,m_i)$ are $NID(0,\sigma^2)$. Define

$$d_k^* = \hat{\beta}_k - \hat{\beta}_{k-1};\quad k = 3,\ldots,M.$$

$$C_k = \sum_{i=1}^{k} m_i (x_i - \bar{x}_k)^2$$

$$\bar{x}_k = \sum_{i=1}^{k} m_i x_{i\cdot}/N_k,\quad N_k = \sum_{i=1}^{k} m_i$$

By R2.2.5 it follows that an analagous test is obtained if $\hat{\alpha}_k - \hat{\alpha}_{k-1}$ is used instead of $\hat{\beta}_k - \hat{\beta}_{k-1}$. Under the null hypothesis that the regression is linear we have

$$E[d_k^*] = 0$$

$$Var[d_k^*] = \sigma^2[C_{k-1}^{-1} - C_k^{-1}]$$

From R2.2.3 it follows that for $k \neq L$; $Cov(d_k^*, d_L^*) = 0$, which means that the d_k^*'s are $NID(0, Var[d_k^*])$.

Define $d_k = d_k^*/[C_{k-1}^{-1} - C_k^{-1}]^{\frac{1}{2}}$, which are $NID(0,\sigma^2)$. If a split-line regression exists the d_k's will have zero mean up to the break-point, but in general have non-zero means thereafter. The mean of the d_k's, $\bar{d}$ say, seems a good statistics to measure any departure from the null hypothesis. Since $\bar{d} \sim N\big(0, \sigma^2/(M-2)\big)$ the test statistic is

$$t = \frac{(M-2)^{\frac{1}{2}}\bar{d}}{s_d},$$

where s_d^2 is an estimate of σ^2, independent of $\bar{d}$. In order to calculate s_d^2, of which there are several possibilities, we need the following result.

R2.3.1

$$\sum_{i=1}^{M} m_i \bar{y}_{i\cdot}^2 = N\bar{y}_{\cdot\cdot}^2 + C_M\hat{\beta}_M^2 + \sum_{k=3}^{M} d_k^2,$$

where $\bar{y}_{i\cdot} = \sum_{j=1}^{m_i} y_{ij}/m_i$, $\bar{y}_{\cdot\cdot} = \sum_{i=1}^{M}\sum_{j=1}^{m_i} y_{ij}/N$

Proof

Define (a) $\underset{\sim}{\bar{y}} = (\sqrt{m_i}\ \bar{y}_{i\cdot})$, which is a vector of M independent normal variates where $\mathrm{Var}(\underset{\sim}{\bar{y}}) = \sigma^2 I_M$

(b) Let $\underset{\sim}{Y}' = (\sqrt{N}\ \bar{y}_{\cdot\cdot}, \sqrt{C_M}\hat{\beta}_M, d_3, \ldots, d_M)$.

We also have $\mathrm{Cov}(\hat{\beta}_\ell, \bar{y}_{\cdot\cdot}) = 0\ \forall\ \ell = 2,..,M$. Since

$$\mathrm{Cov}(\hat{\beta}_\ell, \bar{y}_{\cdot\cdot}) = \sum_{i=1}^{\ell} \sum_{j=1}^{m_i} \sum_{i'=1}^{M} \sum_{j'=1}^{m_{i'}} \frac{(x_i - \bar{x}_\ell)}{C_\ell} \mathrm{Cov}(y_{ij}, y_{i'j'})$$

2.3 (a)

$$= \sigma^2 \sum_{i=1}^{\ell} \sum_{j=1}^{m_i} (x_i - \bar{x}_\ell)/C_\ell = 0$$

By use of R2.2.2 it follows that $\hat{\beta}_M$ is independent of $d_k (\forall k=3,..,M)$ and by use of 2.3(a) it follows that $\hat{\beta}_M$, $\bar{y}_{\cdot\cdot}$ and $d_k (k=3,\ldots,M)$ are all independent normal variates and also $\mathrm{Var}(\underset{\sim}{Y}) = \sigma^2 I_M$.

It is easy to construct an $M \times M$ matrix 'A' such that $\underset{\sim}{Y} = A\underset{\sim}{\bar{y}}$. The matrix A is non-singular since we transform M independent normal variates into M different independent normal variates.

It follows that

$$\sigma^2 I_M = \mathrm{Var}(\underset{\sim}{Y}) = \mathrm{Var}(A\underset{\sim}{\bar{y}}) = A\ \mathrm{Var}(\underset{\sim}{\bar{y}})A' = \sigma^2 AA'$$

$$\Rightarrow AA' = I_M \Rightarrow A'AA' = A' \Rightarrow A'A = I_M.$$

Thus $\underset{\sim}{Y}'\underset{\sim}{Y} = \underset{\sim}{\bar{y}}'\ A'A\underset{\sim}{\bar{y}} = \underset{\sim}{\bar{y}}' I_M \underset{\sim}{\bar{y}} = \underset{\sim}{\bar{y}}'\underset{\sim}{\bar{y}}$

and it follows that

$$N\bar{y}_{\cdot\cdot}^2 + C_M\hat{\beta}_M^2 + \sum_{k=3}^{M} d_k^2 = \sum_{i=1}^{M} m_i \bar{y}_{i\cdot}^2 \quad \text{and since}$$

$$\sum_{k=3}^{M} d_k^2 = \sum_{k=3}^{M} (d_k - \bar{d})^2 + (M-2)\bar{d}^2 \quad \text{the result follows.}$$

$$\sum_{i=1}^{M} m_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = C_M\hat{\beta}_M^2 + (M-2)\bar{d}^2 + \sum_{k=3}^{M} (d_k - \bar{d})^2$$

Thus the test statistic, t, has a student t distribution with the appropriate d.f. depending on the estimate s_d^2.

In estimating σ^2 the following was used;

(a) For $N = M$ the only estimate available is

$s_d^2 = \sum_{k=3}^{M} (d_k - \bar{d})^2/(M-3)$ and then

$t \sim t_{(M-3)}$.

(b) For $N - M$ small, a pooled estimate would be appropriate.

$$s_d^2 = \left(\sum_{k=3}^{m} (d_k - \bar{d})^2 + \sum_{i=1}^{m} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{i.})^2 \right)/(N-3)$$

then $t \sim t_{(N-3)}$

(c) For $N - M$ large then

$$s_d^2 = \sum_{i=1}^{m} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{i.})^2/(N-M)$$

and $t \sim t_{(N-M)}$. This estimate of s_d^2 is reliable, even if the null hypothesis is not true.

An analogous test statistics is $f = t^2 \sim F_{1,K}$; where K is the appropriate degree of freedom. A further test for the non-linearity of the data when $N - M$ is large [using s_d^2 from (c)] is

$$f = \frac{\sum_{k=3}^{M} (d_k - \bar{d})^2/(M-2)}{s_d^2} \sim F_{M-3,N-M}$$

The above procedure in fact generates two tests depending on whether we up-date regression coefficients from the left or right. The result obtained in one test will not depend on the result obtained in the other test. This fact, under certain conditions can give differing test results, depending on whether we update from the left or right. For example suppose that the intersection point of the two regression lines lies to the extreme right of the data. Under such conditions we may find that updating from left to right gives a non-significant test result, whereas from right to left gives a significant test result.

2.4 Test involving a quadratic component.

The purpose of the test developed in 2.3 is to detect any "bend" in the regression line. This can also be done by fitting a curve to the data, and testing the significance of the curve model; compared to the simple linear regression model. A simple and efficient test is to fit by least squares techniques a quadratic regression to the data; the model being

$$E(y_{ij}) = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2; \quad \begin{matrix} i = 1,\ldots,M \\ j = 1,\ldots,m_i \end{matrix}$$

A test for the significance of the quadratic component (α_2) is a test for the existence of a split-line regression model, since this is the only alternative considered.

The test statistic

$$f = \frac{R_Q - R_L}{S_Q} \sim F_{1,N-3}$$

tests for the quadratic component, where

R_Q : Quadratic model regression sum of squares.

R_L : Simple linear regression sum of squares.

S_Q : Mean residual sum of squares for the quadratic model.

2.5 A useful graphical procedure

As in 2.3 this procedure also uses the fact that if a split-line regression model exists, the updated regression coefficients will display certain characteristics. This is due to the fact that as K increases, the data of the second regime begins to affect the regression estimates of the first regression regime.

As in 2.3, two tests are involved here, depending in

which direction we update the regression coefficents.

Define $\hat{\alpha}_k^* = \sum_{i=k-2}^{k+2} \hat{\alpha}_i/5$, $k = 4,\ldots,M-2$ to be a five unit moving average calculated on the updated estimates of α. The only justification for using a moving average is to deflate any large differences between consecutive estimates of α due to some ill-positioning of the data. A plot of $[(\hat{\alpha}_k^*,k);\ k = 4,\ldots,M-2]$ is useful in detecting split-line regression models. Of interest only is the last half of the plot, since if a single regression line exists this should gradually converge to a straight line with zero slope.

If a split-line model exists the plot will exhibit a slope, the magnitude depending on the difference $(\beta_2-\beta_1)$. Hence if a curve is evident in the last half of the plot, it suggests the existence of a split-line regression model.

2.6 Test based on recursive residuals.

The Cusum technique discussed in Brown and Durbin (3) can also be used here. The extension of their procedure to the case of multiply observations at each x value presents no real difficulty.

Define $z_k = \underset{\sim}{l}_k' \underset{\sim}{y}_k^* - m_k \underset{\sim}{x}_k' \underset{\sim}{b}_{k-1}$; $k = 3,\ldots,M$ then

$$\begin{aligned} \mathrm{Var}(z_k) &= \underset{\sim}{l}_k' \mathrm{Var}(\underset{\sim}{y}_k^*) \underset{\sim}{l}_k + m_k^2 \underset{\sim}{x}_k' \mathrm{Var}(\underset{\sim}{b}_{k-1}) \underset{\sim}{x}_k \\ &= \sigma^2 \left(m_k + m_k^2 \underset{\sim}{x}_k' (X_{k-1}' X_{k-1})^{-1} \underset{\sim}{x}_k\right) \\ &= \sigma^2 m_k T_k \end{aligned}$$

Under the null hypothesis of a single regression line it follows that

$$E(z_k) = 0$$

Defining $w_k = z_k (m_k T_k)^{-\frac{1}{2}}$; $k = 3,\ldots,M$ it follows that the w_k's are $NID(0,\sigma^2)$ [see section 2.6.1].

Let $W_k = \sum_{i=3}^{k} w_i / s_w \quad : k = 3,\ldots,M,$

where $s_W^2 = \sum_{i=1}^{M} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}..)^2 - C_M \hat{\beta}_M^2$

$= \sum_{i=1}^{M} \sum_{j=1}^{m_i} (\bar{y}_i. - \bar{y}..)^2 - \sum_{k=3}^{M} d_k^2$

The Cusum technique involves the plotting of the points $[(W_k,k): k = 3,\ldots,M]$. As in 2.5 if a split-line model exists, the means of w_k will be zero up to the break point, but generally non-zero thereafter. The null hypothesis is rejected if the plot of (W_k,k) crosses a set of pre-determined lines. The boundary lines proposed by Brown and Durbin were the lines joining the points $(3, \pm a\sqrt{M-3})$, $(M, \pm 3a\sqrt{M-3})$. The value 'a' was found by solving

$$\Phi(3a) + e^{-4a^2}[1-\Phi(a)] = \alpha/2$$

where $$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{-u^2/2}\, du$$

and α being the significance level of the test.

For $\alpha = 0.05$, $a = 0.948$

$\alpha = 0.01$, $a = 1.143$.

As in 2.5 this procedure generates two tests, depending on the updating procedure.

2.6.1 The relationship between sections 2.3, 2.6

From 2.2.1 it follows that since

$$\underset{\sim}{b}_k - \underset{\sim}{b}_{k-1} = \underset{\sim}{v}_k z_k / T_k$$

we have

P2.6.1 $$d_k^* = \hat{\beta}_k - \hat{\beta}_{k-1} = \frac{(x_k - \bar{x}_{k-1})}{T_k\, C_{k-1}}\, z_k$$

$$= \frac{(x_k - \bar{x}_{k-1})}{C_{k-1}} \left(\frac{m_k}{T_k}\right)^{\frac{1}{2}} w_k \quad , \quad k = 3,\ldots,M-2$$

P2.6.2 $$\Rightarrow \quad w_k = \frac{C_{k-1}}{(x_k - \bar{x}_{k-1})} \left(\frac{T_k}{m_k}\right)^{\frac{1}{2}} (\hat{\beta}_k - \hat{\beta}_{k-1})$$

Since $\mathrm{Var}(z_k) = \sigma^2 m_k T_k$ it follows from P2.6.1 that

$$V(\hat{\beta}_k - \hat{\beta}_{k-1}) = \sigma^2 [C_{k-1}^{-1} - C_k^{-1}]$$

$$= \sigma^2 \frac{(x_k - \bar{x}_{k-1})^2}{C_{k-1}^2} \left(\frac{m_k}{T_k}\right)$$

Thus from P2.6.1 it follows that

P2.6.3 $$d_k = \frac{\sigma(\hat{\beta}_k - \hat{\beta}_{k+1})}{\mathrm{Var}^{\frac{1}{2}}(\hat{\beta}_k - \hat{\beta}_{k-1})} = w_k \; ; \quad k = 3, \ldots, M.$$

Thus the plotting technique of §2.6 is replaced in §2.3 by a test using only the last value W_M of the plotting data.

$$W_M = \sum_{k=3}^{M} w_k / s_w = \sum_{k=3}^{M} d_k / s_w$$

$$= (M-2)\, \bar{d}/s_w$$

$$= \frac{t \times s_d}{s_w}, \quad \text{where } t \text{ is the test statistic}$$

of 2.3.

Numerical examples illustrating the above procedures are given in Chapter 5.

CHAPTER 3

ESTIMATING THE INTERSECTION POINT, γ.

3.1 Introduction

The various techniques employed in estimating γ can be separated into two groups. The first, mentioned only in passing involves a pragmatic approach. This group includes McGee and Careltons (10) use of hierarchical clustering and Wainer's (14) method of approximating first and second derivatives between consecutive data points. The second group are those which attempt to find least squares estimates. We now consider this least squares estimation problem in some detail.

3.2 Method of Least Squares

Let $J_{1k} = \{1,2,\ldots,k\}$, $J_{2k} = \{k+1,\ldots,M\}$ then for $i = 1,2$ define

$$I_{ik} = \{x_j : j \in J_{ik}\}$$

$$C_i(v,w,k) = \sum_{j\in J_{ik}} \sum_{\ell=1}^{m_j} (v_{j\ell}-\bar{v}_{ik})w_{j\ell}$$

$$\bar{v}_{ik} = \sum_{j\in J_{ik}} \sum_{\ell=1}^{m_j} v_{j\ell}/N_{ik}$$

where $N_{ik} = \sum_{j\in J_{ik}} m_j$, $N = N_{1k} + N_{2k}$ and $\eta_k = N_{1k}N_{2k}/N$.

We refer to such a division of the data into two sets as the k'th partition.

Define $\{(\hat{\alpha}_{ik},\hat{\beta}_{ik} : i=1,2),\hat{\sigma}_k^2$ to be the least squares estimates obtained when fitting a pair of lines to the k'th partition.

Then $\hat{\gamma}_k = (\hat{\alpha}_{1k}-\hat{\alpha}_{2k})/(\hat{\beta}_{2k}-\hat{\beta}_{1k})$

where

$$\hat{\beta}_{ik} = C_i(x,y,k)/C_i(x,x,k)$$

3.2.(a) $$\hat{\alpha}_{ik} = \bar{y}_{ik} - \hat{\beta}_{ik}\bar{x}_{ik}$$

$$\hat{\sigma}_k^2 = \sum_{i=1}^{2} [C_i(y,y,k) - \hat{\beta}_{ik} C_i(x,y,k)]/(N-4)$$

Since $\gamma = (\alpha_1-\alpha_2)/(\beta_2-\beta_1)$ we can express the model of chapter 1 as

$$E(y_{ij}) = \begin{cases} \alpha_1 + \beta_1 x_i & , \quad x_i \leqslant \gamma \\ \alpha_1 + \beta_1\gamma + \beta_2(x_i-\gamma), & x_i \geqslant \gamma \end{cases}$$

Then for each partition $\{k : k=2,\ldots,M-2\}$ and for fixed γ (γ=u say), it follows that the linear model is

$$E\begin{bmatrix} \underset{\sim}{y}_{1k} \\ \underset{\sim}{y}_{2k} \end{bmatrix} = X_k \underset{\sim}{\beta}^*_{ku}$$

where $\underset{\sim}{\beta}^{*\prime}_{ku} = (\alpha^*_{1ku}, \beta^*_{1ku}, \beta^*_{2ku})$

$\underset{\sim}{y}' = (\underset{\sim}{y}'_{1k} \cdot \underset{\sim}{y}'_{2k})$

3.2.(b) $\underset{\sim}{y}'_{1k} = (\underset{\sim}{y}'_1 \cdot \ldots \underset{\sim}{y}'_k)$, $\underset{\sim}{y}'_{2k} = (\underset{\sim}{y}_{k+1} \cdot \ldots \cdot \underset{\sim}{y}'_M)$

$\underset{\sim}{x}'_n = x_n \underset{\sim}{1}_{m_n}$ since $x_{n,j} = x_n$ for $j = 1,\ldots,m_n$

3.2.(c) Define $\underset{\sim}{\hat{\beta}}^{*\prime}_{ku} = (\hat{\alpha}^*_{1ku}, \underset{\sim}{\hat{\beta}}^*_{1ku}, \underset{\sim}{\hat{\beta}}^*_{2ku})$ as the least squares estimate of $\underset{\sim}{\beta}^{*\prime}_{ku}$ obtained by fitting a pair of regression lines, meeting at $x = u$, to the k'th partition and let

$$Z_k = [X_k : \underset{\sim}{y}]$$

3.2.(d)
$$= \begin{bmatrix} \underset{\sim}{1}_{N_{1k}} & \underset{\sim}{x}_{1k} & \underset{\sim}{0}_{N_{1k}} & \vdots & \underset{\sim}{y}_{1k} \\ \underset{\sim}{1}_{N_{2k}} & u\underset{\sim}{1}_{N_{2k}} & \underset{\sim}{X}_{k-u}\underset{\sim}{1}_{N_{2k}} & \vdots & \underset{\sim}{y}_{2k} \end{bmatrix}.$$

By applying the sweep operator [Dempster (4), p.62-65] to the symmetric matrix $Z'_k Z_k$ using as successive pivots the first three diagonal elements, the fourth diagonal element is changed to the residual sum of squares which we derive as a function of u, given by

3.2.(e) $$S_k^2(u) = C_1(y,y,M) - \frac{[a_k^2(u) f_k(u) + b_k(u) c_k^2(u) - 2c_k(u) g_k(u) a_k(u)]}{e_k(u)}$$

where

$$a_k(u) = C_1(x,y,k) + \eta_k(\bar{y}_{1k}-\bar{y}_{2k})(\bar{x}_{1k}-u)$$

$$b_k(u) = C_1(x,x,k) + \eta_k(\bar{x}_{1k}-u)^2$$

$$c_k(u) = C_2(x,y,k) - \eta_k(\bar{y}_{1k}-\bar{y}_{2k})(\bar{x}_{2k}-u)$$

$$g_k(u) = -\eta_k(\bar{x}_{2k}-u)(\bar{x}_{1k}-u)$$

$$f_k(u) = C_2(x,x,k) + \eta_k(\bar{x}_{2k}-u)^2$$

and $e_k(u) = C_k - 2D_k u + E_k u^2$, *with* C_k *now being defined as*

$$C_k = C_1(x,x,k)C_2(x,x,k) + \eta_k[\bar{x}_{1k}^2 C_2(x,x,k) + \bar{x}_{2k}^2 C_1(x,x,k)],$$

and $D_k = \eta_k[\bar{x}_{1k}C_2(x,x,k) + \bar{x}_{2k}C_1(x,x,k)]$

$$E_k = \eta_k[C_1(x,x,k) + C_2(x,x,k)]$$

The equation of 3.2.(e) can be rewritten as

3.2.(f)
$$S_k^2(u) = S_0^2 - \frac{(\hat{\beta}_{2k}-\hat{\beta}_{1k})^2[C_k - D_k(\hat{\gamma}_k+u) + E_k\hat{\gamma}_k u]^2}{C_1(x,x,M)[C_k - 2D_k u + E_k u^2]} \quad \text{(Hinkley)}$$

where

$S_0^2 = C_1(y,y,M) - C_1(x,x,M).\hat{\beta}_{1M}^2$, the residual sum of squares for a single regression line fitted to all the data.

The estimates of $\underset{\sim}{\beta}^*_{ku}$ can also be written as functions of u,

$$\hat{\beta}^*_{1ku} = [f_k(u)a_k(u) - g_k(u)c_k(u)]/e_k(u)$$

3.2.(g)
$$\hat{\beta}^*_{2ku} = [b_k(u)c_k(u) - g_k(u)a_k(u)]/e_k(u)$$

$$\hat{\alpha}^*_{1ku} = \bar{y}_{1k} - \hat{\beta}^*_{1ku}\,\bar{x}_{1k} + \frac{N_{2k}}{N}[(\bar{y}_{2k}-\bar{y}_{1k}) + \hat{\beta}^*_{1ku}(\bar{x}_{1k}-u) - \hat{\beta}_{2ku}(\bar{x}_{2k}-u)]$$

Define $S^2(u) = S_k^2(u)$; $x_k \leqslant u \leqslant x_{k+1}$, $k = 2,\ldots,M-2$.

Then since

(a) $S_k^2(u)$ is a continuous function

and (b) $S_{k-1}^2(x_k) = S_k^2(x_k)\ \forall\ k$ [Hinkley],

it follows that $S^2(u)$ is also a continuous function.

Under the assumptions of the model, the least squares estimate of γ for each partition (k say) must lie in the

interval $[x_k, x_{k+1}]$. It follows that the exhaustive set of possible solutions for the least squares estimate, $\hat{\gamma}$, of γ is given by

$$T = \{[X],[G]\} \quad \text{where}$$

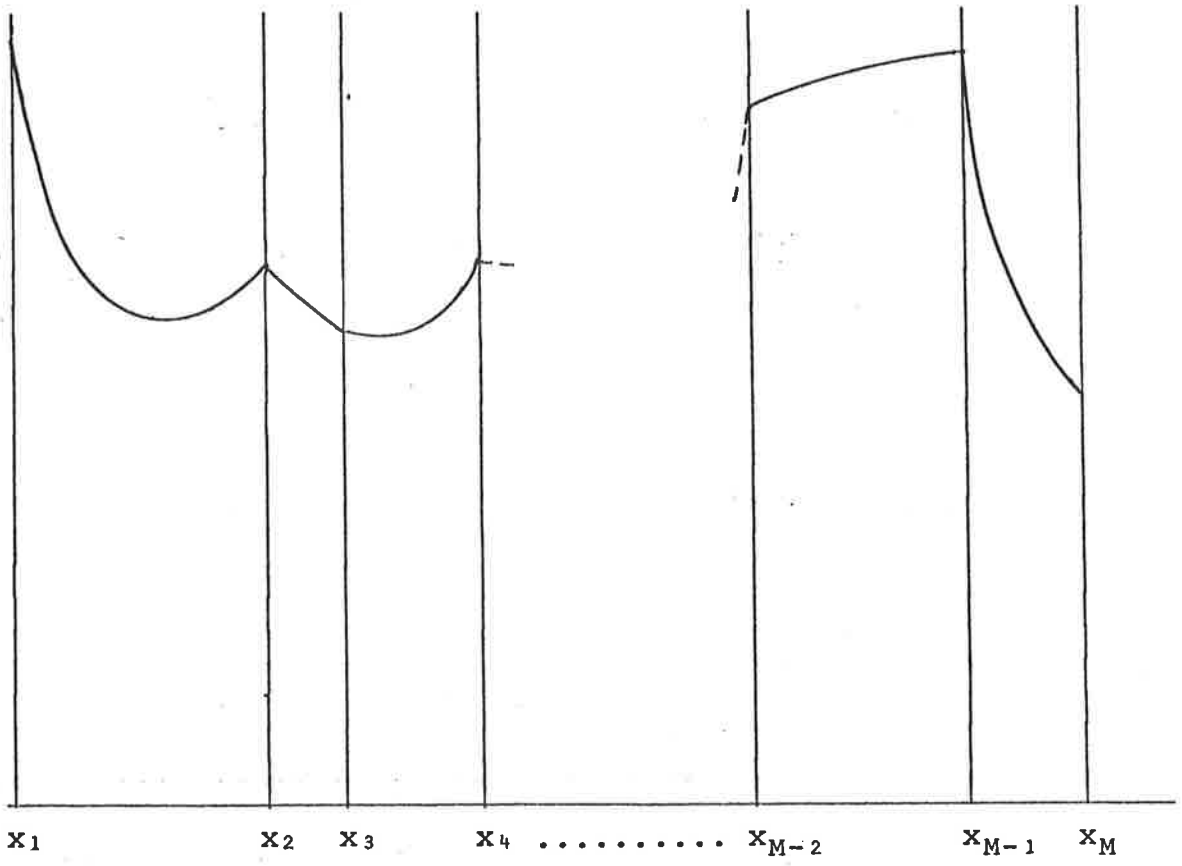
$$G = \{\hat{\gamma}_k; k=2,\ldots,M-2\} \quad \text{and}$$

as before $$X = \{x_i : i=1,\ldots,M-1\}.$$

Then

$$S^2(\hat{\gamma}) = \underset{\text{all } k}{\text{Min}}\ S^2(u) = \underset{\text{all } k}{\text{Min}}\ \{\text{Min}\ S_k^2(u); x_k \leqslant u \leqslant x_{k+1}\}.$$

A possible form for $S^2(u)$ is given below,



3.3. In this section, which is an extension of Hinkley's work, we consider some properties of $S_k^2(u)$.

Let

1. $\hat{\xi}_k = (\hat{\beta}_{2k}-\hat{\beta}_{1k})^2/C_1(x,x,M)$

2. $M_k = \hat{\xi}_k(C_kE_k-D_k^2) > 0$ since

$$C_kE_k-D_k^2 = \eta_kC_1(x,x,k)C_2(x,x,k)[C_1(x,x,k)+C_2(x,x,k)] + \eta_k^2(\bar{x}_{1k}-\bar{x}_{2k})^2\, C_1(x,x,k)C_2(x,x,k) > 0$$

3. For each u,k there exists a τ such that

3.3.(a) $u = \hat{\gamma}_k + \tau$.

4. $e_k(u) = C_k - 2D_ku + E_ku^2$

$$= E_k(u-D_k/E_k)^2 + (C_k-D_k^2/E_k) > 0 \quad \text{for all } u.$$

5. $Z_k^2(u) = \hat{\xi}_k[C_k-D_k(\hat{\gamma}_k+u) + E_k\hat{\gamma}_ku]^2/e_k(u)$

$$= Z_k^2(\hat{\gamma}_k) - \tau^2M_k/e_k(u)$$

Then 3.2.(f) can be expressed as

$$S_k^2(u) = S_0^2 - Z_k^2(u) = S_0^2 - Z_k^2(\hat{\gamma}_k) + \tau^2M_k/e_k(u)$$

3.3.1. The turning points of $S_k^2(u)$

$$\frac{dS_k^2(u)}{du} = 2M_k(u-\hat{\gamma}_k)(\delta_k-u)(D_k-E_k\hat{\gamma}_k)$$

where

$$\delta_k = (C_k-D_k\hat{\gamma}_k)/(D_k-E_k\hat{\gamma}_k)$$

3.3.(b) $$= \hat{\gamma}_k + e_k(\hat{\gamma}_k)/(D_k-E_k\hat{\gamma}_k)$$

$$= D_k/E_k + (C_k-D_k^2/E_k)/(D_k-E_k\hat{\gamma}_k)$$

It follows that $S_k^2(u)$ has a Minimum turning point at $u = \hat{\gamma}_k$ and a Maximum turning point when $u = \delta_k$.

By use of 3.3.(b) we have

$$S_k^2(\delta_k) = S_0^2 - Z_k^2(\hat{\gamma}_k) + (\delta_k-\hat{\gamma}_k)^2M_k/e_k(\delta_k)$$

$$= S_0^2 - Z_k^2(\hat{\gamma}_k) + \hat{\xi}_k\, e_k(\hat{\gamma}_k)$$

$$= S_0^2 \quad \text{since} \quad Z_k^2(\hat{\gamma}_k) = \hat{\xi}_ke_k(\hat{\gamma}_k)$$

If for each partition of the data we construct lines to pass through the point $u = \delta_k$, we in fact obtain the single

linear regression line fitted to all the data. This is easily verified by putting $u = \delta_k$ into the equations of 3.2.(g) to obtain

$$\hat{\beta}^*_{iku} = \hat{\beta}_{ik} \qquad i = 1,2$$

$$\hat{\alpha}^*_{iku} = \hat{\alpha}_{ik}.$$

Define δ^*_k, δ_k respectively as the maximum turning points of $S^2_k(u)$, depending on whether $\hat{\gamma}_k > D_k/E_k$ or $\hat{\gamma}_k < D_k/E_k$.

Since $\lim_{u\to\pm\infty} \tau^2/e_k(u) = E_k^{-1}$ and

$$Z^2_k(\hat{\gamma}_k) = \frac{M_k(\hat{\gamma}_k - D_k/E_k)^2}{(C_k - D_k^2/E_k)} + \frac{M_k}{E_k} > \frac{M_k}{E_k}$$

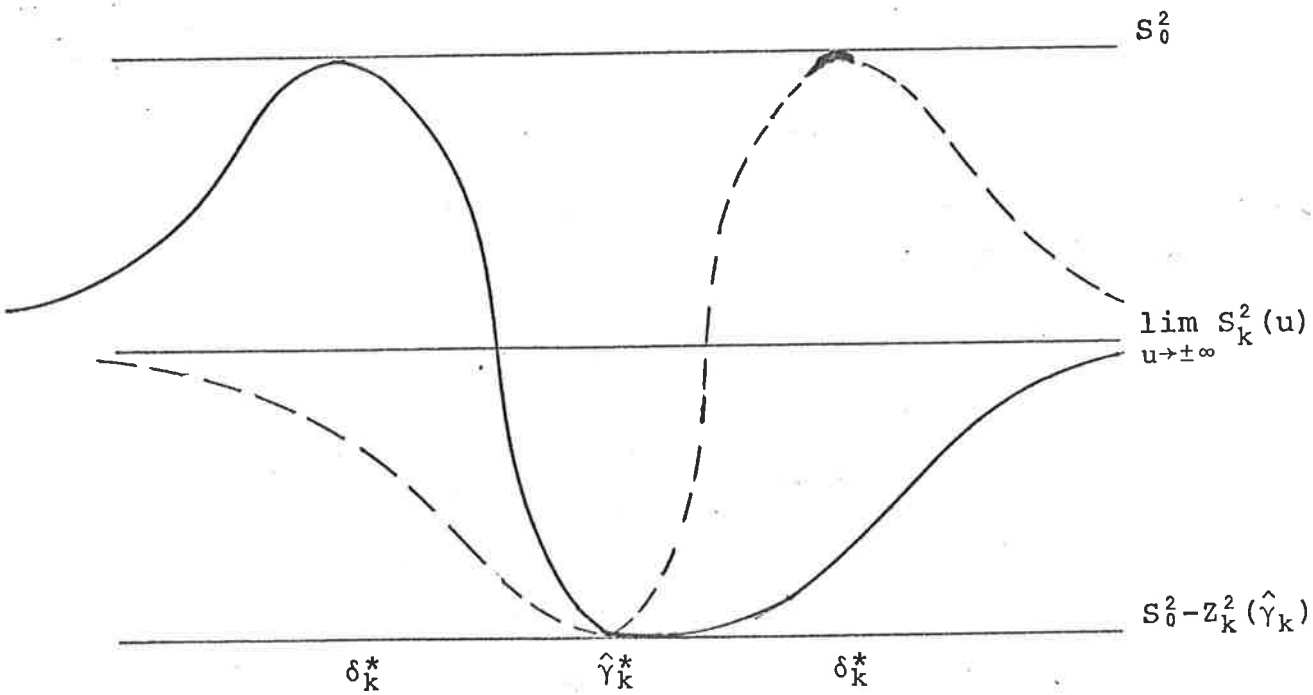
it follows that

3.3.(c)

$$S^2_0 - Z^2_k(\hat{\gamma}_k) < \lim_{\mu\to\pm\infty} S^2_k(u) = S^2_0 - \frac{M_k(\hat{\gamma}_k - D_k/E_k)^2}{(C_k - D_k^2/E_k)} < S^2_0$$

As a consistency check it can be shown that, for each partition of the data, $\lim_{\mu\to\pm\infty} S^2_k(u)$ is the residual sum of squares obtained by fitting parallel lines to the two sets of data.

D3.3.1 The shape of $S^2_k(u)$.



It follows that $S_k^2(u)$ is a monotone increasing function as $|\hat{\gamma}_k - u|$ increases, so long as

(a) $u > \delta_k^*$ for $\hat{\gamma}_k > D_k/E_k$

3.3.(d)

(b) $u < \delta_k$ for $\hat{\gamma}_k < D_k/E_k$

3.3.2 The function $S_\ell^2(u)$ as $\hat{\beta}_{2\ell}$ approaches $\hat{\beta}_{1\ell}$, where the least squares estimate of γ is such that $x_\ell \leqslant \hat{\gamma} < x_{\ell+1}$.

Define $\eta = (\beta_2 - \beta_1)/\sigma$, then the least squares estimate of B is given by $\hat{B} = (\hat{\beta}_2 - \hat{\beta}_1)/\sigma$ (if σ unknown, replace it by $\hat{\sigma}$); where $\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}$ are the least squares estimates of β_1, β_2, σ. Asymptotically $\hat{B}$ has a normal distribution with mean B and variance $\left(C_1^{-1}(x,x,T) + C_2^{-1}(x,x,T)\right)$, where $x_T \leqslant \gamma < x_{T+1}$. Hinkley has shown empirically that in finite samples, a positive bias in $\hat{B}$ must be expected because the minimization of $S^2(u)$ is associated with maximizing the difference between the two regression slopes. Furthermore the estimation problem becomes unreliable for certain combinations of B, γ and N. These ill defined cases generally occur in situations when $|B|.\eta_T < 5$ approximately. It follows that for fixed σ the positive bias in $\hat{B}$ and the chance of an ill defined case will increase as β_2 approaches β_1. This can be demonstrated by a study of the shape of $S_\ell^2(u)$ as $\hat{\beta}_{2\ell} \to \hat{\beta}_{1\ell}$. Although $\hat{\beta}_{1\ell}, \hat{\beta}_{2\ell}$ are constants, we can consider $\hat{\beta}_{2\ell} \to \hat{\beta}_{1\ell}$ if we fix the x values and manipulate the y values such that α_1, α_2 remain fixed, but β_1, β_2 are allowed to vary. Under such conditions $S_\ell^2(u)$ also becomes a function of $\hat{\beta}_{2\ell}, \hat{\beta}_{1\ell}$ and we can express 3.2.(f) as

$$S_\ell^2(u) = S_0^2 - \frac{[(\hat{\beta}_{2\ell}-\hat{\beta}_{1\ell})(C_\ell-D_\ell u) + (\hat{\alpha}_{1\ell}-\hat{\alpha}_{2\ell})(E_\ell u-D_\ell)]^2}{e_\ell(u)\ C_1(x,x,M)}$$

then

$$\lim_{\hat{\beta}_{2\ell}\to\hat{\beta}_{1\ell}} S_\ell^2(u) = S_0^2 - \frac{(\hat{\alpha}_{1\ell}-\hat{\alpha}_{2\ell})^2(E_\ell u-D_\ell)^2}{e_\ell(u)\ C_1(x,x,M)}$$

$$= S_0^2 - \frac{(\hat{\alpha}_{1\ell}-\hat{\alpha}_{2\ell})^2 E_\ell}{C_1(x,x,M)}\left[\frac{(D_\ell-E_\ell u)^2}{(D_\ell-E_\ell u)^2+(C_\ell E_\ell-D_\ell^2)}\right]$$

has the following properties:

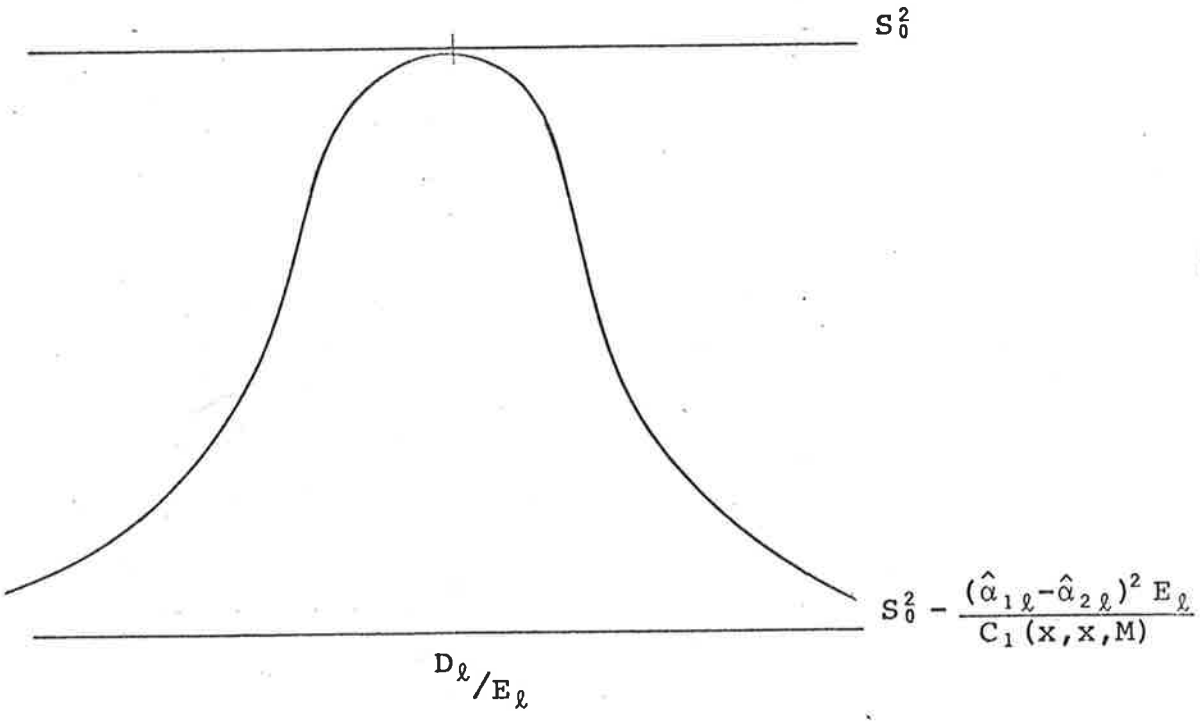
(a) monotone decreasing as $|u-D_\ell/E_\ell|$ increases

(b) maximum value of S_0^2 when $u = D_\ell/E_\ell$.

(c) a lower asymptote of

$$S_0^2 - \frac{(\hat{\alpha}_{1\ell}-\hat{\alpha}_{2\ell})^2 E_\ell}{C_1(x,x,M)}$$

Thus as $\hat{\beta}_{2\ell} \to \hat{\beta}_{1\ell}$ the assumption that $\beta_2 \neq \beta_1$ is apparently violated with the form of $S_\ell^2(u)$ changing from that of D3.3.1 to the following



We now consider some previous approaches for the estimation of the intersection point γ.

3.4 Hudson's (8) method

The following theorem is used by Hudson to construct an interval in which $S_k^2(u)$ is an increasing function as $|\tau|$ increases.

McLaren's Theorem (8)

"Let $\tau = \hat{\gamma}_k - u$, then if $\hat{\gamma}_k$ lies within the interval $[\bar{x}_{1k}, \bar{x}_{2k}]$, $S_k^2(u)$ is a monotone increasing function as $|\tau|$ increases; so long as u is contained in $[\bar{x}_{1k}, \bar{x}_{2k}]$." We shall now demonstrate a simple proof for this theorem.

Since $\hat{\gamma}_k$ lies within $[\bar{x}_{1k}, \bar{x}_{2k}]$ it follows that

$$\bar{x}_{1k} - \hat{\gamma}_k < 0 \quad \text{and} \quad \bar{x}_{2k} - \bar{x}_{1k} > 0$$

then for $D_k - E_k\hat{\gamma}_k > 0$ it can be readily shown that $\delta_k > \bar{x}_{2k}$, for example since $\bar{x}_{1k} - \hat{\gamma}_k < 0$ it follows that

$$C_1(x,x,k) > \eta_k(\bar{x}_{2k}-\bar{x}_{1k})(\bar{x}_{1k}-\hat{\gamma}_k) \quad \text{for all } k.$$

$$\Leftrightarrow C_1(x,x,k).C_2(x,x,k) > \eta_k(\bar{x}_{2k}-\bar{x}_{1k})(\bar{x}_{1k}-\hat{\gamma}_k)C_2(x,x,k)$$

and by adding $\eta_k\bar{x}_{2k}(\bar{x}_{2k}-\hat{\gamma}_k)C_1(x,x,k)$ to each side, the required result is obtained. Similarly for $D_k - E_k\hat{\gamma}_k < 0$ we have $\delta_k^* < \bar{x}_{1k}$.

It follows that for $\hat{\gamma}_k$ in the interval $[\bar{x}_{1k}, \bar{x}_{2k}]$ we have

$$\delta_k^* < \bar{x}_{1k} \leqslant \hat{\gamma}_k \leqslant \bar{x}_{2k} < \delta_k$$

and by use of 3.3.(d) the theorem is proved.

The searching procedure used to find the least squares estimate is

1. (i) $K \leftarrow 1$

   (ii) $U \leftarrow S_0^2$

   (iii) $v \leftarrow x_1$

2. (i) $k \leftarrow k + 1$

   (ii) if $k > M$ $(M > 3)$ stop.

   (iii) find $\hat{\gamma}_k$ (3.2.(a)), $S_k^2(\hat{\gamma}_k)$ (3.2.(e)).

   (iv) if $U \leqslant S_k^2(\hat{\gamma}_k)$ go to 2

   (v) if $\hat{\gamma}_k$ lies outside $[x_k, x_{k+1}]$ go to 3

   (vi) $U \leftarrow S_k^2(\hat{\gamma}_k)$, $v \leftarrow \hat{\gamma}_k$

   (vii) go to 2.

3. (i) if $\hat{\gamma}_k$ lies outside $[\bar{x}_{1k}, \bar{x}_{2k}]$ go to 4.

   (ii) for $\hat{\gamma}_k < x_k$, $U_0 \leftarrow S_k^2(x_k)$, $v_0 \leftarrow x_k$

   (iii) for $\hat{\gamma}_k > x_{k+1}$, $U_0 \leftarrow S_k^2(x_{k+1})$, $v_0 \leftarrow x_{k+1}$

   (iv) if $U \leqslant U_0$ go to 2.

   (v) $U \leftarrow U_0$, $v \leftarrow v_0$

   (vi) go to 2.

4. (i) $U_0 = \text{Min}\{S_k^2(x_k), S_k^2(x_{k+1})\} = S_k^2(v_0)$

   (ii) if $U \leqslant U_0$ go to 2

   (iii) $U \leftarrow U_0$, $v \leftarrow v_0$

   (iv) go to 2.

The final value of v is the least squares estimate, $\hat{\gamma}$, of γ with $U = S^2(\hat{\gamma})$. The least squares estimates of the regression parameters can be easily found by use of 3.2.(g), and also $\hat{\sigma}^2 = U/(N-4)$.

3.5 Hinkley's (7) method

A slight improvement on Hudson's procedure with efficiencies being achieved by

(a) the use of the whole range of u (see 3.3.(d)) in which $S_k^2(u)$ increases as $|\tau|$ increases. For example if $u < \delta_k$ $(D_k - E_k\hat{\gamma}_k > 0)$ Hinkley's procedure requires only one step, whereas Hudson's would require two steps if u lies outside the interval $[\bar{x}_{1k}, \bar{x}_{2k}]$.

(b) simplifying $S_k^2(u)$ to $S_0^2 - Z_k^2(u)$ and finding the least squares estimate of γ by maximizing the function $Z_k^2(u)$.

The procedure is as follows

1. (i) $K \leftarrow 1$
   (ii) $U \leftarrow 0$
   (iii) $v \leftarrow x_1$

2. (i) $k \leftarrow k + 1$
   (ii) if $k > M$ $(M > 3)$ stop
   (iii) find $\hat{\gamma}_k$ (3.2.(a)), $Z_k^2(\hat{\gamma}_k)$ (3.3.(a)).
   (iv) if $U \geqslant Z_k^2(\hat{\gamma}_k)$ go to 2
   (v) if $\hat{\gamma}_k$ lies outside $[x_k, x_{k+1}]$ go to 3
   (vi) $U \leftarrow Z_k^2(\hat{\gamma}_k)$, $v \leftarrow \hat{\gamma}_k$
   (vii) go to 2.

3. (i) If $D_k - E_k\hat{\gamma}_k < 0$ and $x_k < \delta_k^*$ go to 4
   (ii) If $D_k - E_k\hat{\gamma}_k > 0$ and $x_{k+1} > \delta_k$ go to 4
   (iii) for $\hat{\gamma}_k < x_k$, $U_0 \leftarrow Z_k^2(x_k)$, $v_0 \leftarrow x_k$
   (iv) for $\hat{\gamma}_k > x_{k+1}$, $U_0 \leftarrow Z_k^2(x_{k+1})$, $v_0 \leftarrow x_{k+1}$
   (v) if $U \geqslant U_0$ go to 2
   (vi) $U \leftarrow U_0$, $v \leftarrow v_0$ go to 2.

4. (i) $U_0 = \text{Max}\{Z_k^2(x_k), Z_k^2(x_{k+1})\} = S_k^2(v_0)$

(ii) if $U \geq U_0$ go to 2

(iii) $U \leftarrow U_0$, $v \leftarrow v_0$ go to 2.

The final value of v is the least squares estimate of γ and $\hat{\sigma}^2 = (S_0-U)/(N-4)$. The least squares estimates of the regression parameters are found as in §3.4.

3.6 A Refinement of Hinkley's Method.

For this approach, suggested by W.N. Venables, we start with a close upper bound to the minimum residual sum of squares. We then consider, in sequence, each possible range $\{[x_k, x_{k+1}],\ k=2,\ldots,M-2\}$ in which the least squares estimate of γ could lie. For each partition the calculations are stopped as soon as it is clear that $\hat{\gamma}$ could not possibly lie within the range, otherwise the upper-bound is revised and we continue with the searching procedure for the remaining partitions.

An algorithm for least squares estimation of a split-line model.

Denote S_μ^2 as the upper bound for the minimum residual sum of squares.

1. Calculating an initial upper bound.

   One method is to use a sub-interval (x_{ℓ_1}, x_{ℓ_2}) of the data, with ℓ_1, ℓ_2 given as

   $\ell_1 = [M/3]$, $\ell_2 = [2M/3]$: with [] denoting the integer part. A study of a data plot could be more informative in that we may be able to choose ℓ_1, ℓ_2 such that (x_{ℓ_1}, x_{ℓ_2}) covers the estimated intersection point.

   Let $L = [(\ell_1+\ell_2)/2]$ and fit regression lines to the partitions I_{L1}, I_{L2}; giving $\hat{\gamma}_L$.

   Then let

   $$Z_\mu^2 \leftarrow Z_L^2(\hat{\gamma}_L) \quad \text{if} \quad x_L \leqslant \hat{\gamma}_L \leqslant x_{L+1}$$

   $$Z_\mu^2 \leftarrow \text{Max}\{Z_L^2(x_L), Z_L^2(x_{L+1})\} = Z_L^2(w)$$

   $$\text{if} \quad \hat{\gamma}_L < x_L, \text{ or } \hat{\gamma}_L > x_{L+1}$$

and it follows that

$$S_\mu^2 = S_0^2 - Z_\mu^2$$

2. Searching Procedure.

Using Z_μ^2 we employ Hinkley's procedure of 3.5, except that step 1 is changed to the following.

(i) $k \leftarrow 1$

(ii) $U \leftarrow Z_\mu^2$

(iii) $v \leftarrow \hat{\gamma}_L$ if $x_L \leqslant \hat{\gamma}_L \leqslant x_{L+1}$

or $v \leftarrow w$ (see previous page)

As before, the final value of v will be the least squares estimate of γ with $\hat{\sigma}^2 = (S_0-U)/(N-4)$. The least squares regression parameter estimates are found in the usual way.

3.7 An Interval Estimation Approach.

This procedure aims at finding an interval in which the least squares estimate of γ is contained. Then by applying Hinkley's method only to the values within this interval we obtain an estimate for γ. Since only a subset of X is searched a considerable saving in computational time can be achieved. The disadvantage of this method is that no guarantee can be given that the optimum least squares estimate of γ is contained within this interval, yet this proved to be of no disadvantage in the variety of data sets analysed.

The method is as follows:

1. Finding an initial interval (a^0,b^0).

Using the procedure of method 3.6 part 1 we obtain $\hat{\gamma}_L$, then set

$$\hat{\gamma}^0 = \begin{cases} \hat{\gamma}_L & \text{if } x_2 < \hat{\gamma}_L < x_{M-1} \\ (x_{\ell_1}+x_{\ell_2})/2 & \text{if } \hat{\gamma}_L \leq x_2 \text{ or } \hat{\gamma}_L \geq x_{M-1} \end{cases}$$

and let

$$a^0 = \hat{\gamma}^0 - r, \quad b^0 = \hat{\gamma}^0 + r \quad \text{where}$$

$$2r = \text{Min}\{\hat{\gamma}^0 - x_1,\ x_M - \hat{\gamma}^0\}.$$

2. Updating the initial interval to (a^1, b^1).

Define $\mu_\ell = \alpha_\ell + \beta_\ell x$; $\ell = 1,2$ and

$$E(y/x) = p(x)\mu_1 + q(x)\mu_2$$

3.7(a)
$$= p(x)\ (\alpha_1+\beta_1 x) + q(x)\ (\alpha_2+\beta_2 x)$$

where $q(x) = 1 - p(x)$,

the probability $p(x)$ being defined as

$$p(x) = \begin{cases} 1 & x \leq a^0 \\ F(x) & a^0 \leq x \leq b^0 \\ 0 & x \geq b^0 \end{cases}$$

where $F(x)$ is the cumulative logistic distribution function given by

$F(x) = 1 - \{1+e^{-(x-\xi_0)/\xi_1}\}^{-1}$; $\xi_1 > 0, |\xi_0| < \infty$ where ξ_0, ξ_1 are given the values

$\xi_0 = (a^0+b^0)/2$

$\xi_1 = (b^0-a^0)/20$, so that

(i) $F(a^0) \doteqdot 1$ (= 0.999954)

(ii) $F(b^0) \doteqdot 0$ (0.000046)

(iii) $\lim_{x \to a^0_-} \frac{d}{dx} E(y/x) = \beta_1 \doteqdot \lim_{x \to a^0_+} \frac{d}{dx} E(y/x)$

(iv) $\lim_{x \to b^0_+} \frac{d}{dx} E(y/x) = \beta_2 \doteqdot \lim_{x \to b^0_-} \frac{d}{dx} E(y/x)$.

where

$$\frac{d}{dx} E(y/x) = \{\alpha_1-\alpha_2+(\beta_1-\beta_2)x\}\frac{dp(x)}{dx} + p(x)(\beta_1-\beta_2) + \beta_2$$

For the data

$\{(x_i, y_{ij}), j=1,\ldots,m_i;\ i=1,\ldots,M\}$

we now replace the split line model by the smooth transition function model of 3.7(a). Denoting

$\underset{\sim}{y} = (y_{ij})$, $\underset{\sim}{p}' = \left(p(x_1)\underset{\sim}{1}'_{M_1}. \ \ldots \ p(x_m)\underset{\sim}{1}'_{m_M}\right)$;

then $E(\underset{\sim}{y}/\underset{\sim}{x}) = X_p\ \underset{\sim}{\beta}_p$,

where $X_p = [\underset{\sim}{p}, p\underset{\sim}{x}, \underset{\sim}{q}, q\underset{\sim}{x}]$; it then follows from least squares theory that

$\hat{\underset{\sim}{\beta}}_p = (X_p'X_p)^{-1}\ X_p'\underset{\sim}{y}$, where we denote

$\hat{\underset{\sim}{\beta}}_p' = (\hat{\alpha}_{1p}, \hat{\beta}_{1p}, \hat{\alpha}_{2p}, \hat{\beta}_{2p})$

and $\hat{\gamma}_p = (\hat{\alpha}_{1p} - \hat{\alpha}_{2p})/(\hat{\beta}_{2p} - \hat{\beta}_{1p})$.

Let $\delta^0 = b^0 - a^0$, then the updated interval is given by

$a^1 = \hat{\gamma}_p - \delta^0/3$

$b^1 = \hat{\gamma}_p + \delta^0/3$

3. Finding the estimate of γ

Hinkley's search procedure is then applied to the interval (a^1, b^1) and the estimate of γ giving the minimum residual sum of squares within this interval is found. In all cases studied this estimate proved to be the least squares estimate of γ.

CHAPTER 4

INFERENCE

In this chapter we use the notation $\{(x_i,y_i);i=1,\ldots N\}$ to label a data set containing N points.

Let $\hat{\gamma}_N$ denote the least squares estimate of γ and suppose $x_{t_N} \leqslant \hat{\gamma}_N < x_{t_N+1}$. Define $(\hat{\alpha}^*_{it_N},\hat{\beta}^*_{it_N};i=1,2)$ as the least squares estimates for the parameters of separate regression lines fitted to either side of the partition;

$$\{(x_1,y_1),\ldots(x_{t_N},y_{t_N});\ (x_{t_N+1},y_{t_N+1}),\ldots(x_N,y_N)\}.$$

Let $\hat{\gamma}^*_{t_N} = (\hat{\alpha}^*_{1t_N}-\hat{\alpha}^*_{2t_N})/(\hat{\beta}^*_{2t_N}-\hat{\beta}^*_{1t_N})$. It follows that

$\hat{\gamma}^*_{t_N} = \hat{\gamma}_N$ if $x_{t_N} < \hat{\gamma}_N < x_{t_N+1}$

or $\hat{\gamma}^*_{t_N}$ lies outside the interval

$[x_{t_N},x_{t_N+1}]$ if $\hat{\gamma}_N = x_{t_N}$.

Feder and Sylvester (6) say that by deleting relatively few observations from the data, classical techniques can be used to derive the asymptotic distribution of $\hat{\gamma}_N$. Hinkley has shown asymptotic normality for $\hat{\gamma}_N$ in the case of equally spaced x values. He further indicates that this holds in more general cases with the asymptotic normality of $\hat{\gamma}_N$ dependent on the sequence of configurations of the x values.

We now consider the hypothetical case where $x_{T_N} \leqslant \gamma < x_{T_N+1}$ and T_N is supposed known.

4.1 Large sample variance of $\hat{\gamma}^*_{T_N}$.

Let $\hat{\alpha}^*_{T_N} = \hat{\alpha}^*_{1T_N} - \hat{\alpha}^*_{2T_N}$ and $\hat{\beta}^*_{T_N} = \hat{\beta}^*_{2T_N} - \hat{\beta}^*_{1T_N}$;

then

$$\mathrm{Var}(\hat{\alpha}^*_{T_N}) = \sigma^2\left(\frac{1}{T_N}+\frac{1}{N-T_N}+\frac{\bar{X}^2_{1T_N}}{S_{1T_N}}+\frac{\bar{X}^2_{2T_N}}{S_{2T_N}}\right)$$

$$\mathrm{Var}(\hat{\beta}^*_{T_N}) = \sigma^2\left(\frac{1}{S_{1T_N}} + \frac{1}{S_{2T_N}}\right) \quad \text{and}$$

$$\mathrm{Cov}(\hat{\alpha}^*_{T_N}, \hat{\beta}^*_{T_N}) = \sigma^2\left(\frac{\bar{X}_{1T_N}}{S_{1T_N}} + \frac{\bar{X}_{2T_N}}{S_{2T_N}}\right), \quad \text{where}$$

$$\bar{X}_{1T_N} = \sum_{j=1}^{T_N} x_j/T_N, \quad S_{1T_N} = \sum_{j=1}^{T_N}(x_j - \bar{X}_{1T_N})^2$$

and similarly for $\bar{X}_{2T_N}$ and S_{2T_N}.

Since $\hat{\gamma}^*_{T_N}$ is the ratio of two normal variates the variance of $\hat{\gamma}^*_{T_N}$ does not exist, but the large sample distribution of $\hat{\gamma}^*_{T_N}$ can be approximated by a distribution which has a mean γ,

and variance,

$$4.1(a)\quad V(\hat{\gamma}^*_{T_N}) = \left(\frac{\partial\hat{\gamma}^*_{T_N}}{\partial\hat{\alpha}^*_{T_N}}\right)^2 \mathrm{Var}(\hat{\alpha}^*_{T_N}) + \left(\frac{\partial\hat{\gamma}^*_{T_N}}{\partial\hat{\beta}^*_{T_N}}\right)^2 \mathrm{Var}(\hat{\beta}^*_{T_N})$$

$$+ 2\left(\frac{\partial\hat{\gamma}^*_{T_N}}{\partial\hat{\alpha}^*_{T_N}}\right)\left(\frac{\partial\hat{\gamma}^*_{T_N}}{\partial\hat{\beta}^*_{T_N}}\right) \mathrm{Cov}(\hat{\alpha}^*_{T_N}, \hat{\beta}^*_{T_N}), \quad \text{where}$$

the partial derivatives are evaluated at $\alpha^*(=\alpha_1-\alpha_2)$ and $\beta^*(=\beta_2-\beta_1 : \beta_1 \neq \beta_2)$, giving

$$V(\hat{\gamma}^*_{T_N}) = \frac{\sigma^2}{(\beta_2-\beta_1)^2}\left\{\frac{1}{T_N} + \frac{1}{N-T_N} + \frac{(\gamma-\bar{X}_{1T_N})^2}{S_{1T_N}} + \frac{(\gamma-\bar{X}_{2T_N})^2}{S_{2T_N}}\right\}.$$

Asymptotically, whether $\hat{\gamma}^*_{T_N}$ converges to γ will depend on the sequence of configurations of the x values. Only for those sequences where T_N, $N-T_N$, S_{1T_N} and S_{2T_N} all diverge will $V(\hat{\gamma}^*_{T_N})$ converge to zero and it then follows that $\hat{\gamma}^*_{T_N}$ will be a consistent estimator of γ.

4.2 Large sample variance of $\hat{\gamma}^*_{T_N}$ for equally spaced x values.

Let $\delta = x_i - x_{i-1}$ for $i = 2,\ldots,N$ and we consider firstly the expression $(\gamma-\bar{X}_{1T_N})^2/S_{1T_N}$. Since $x_k = x_1+(k-1)\delta$ for all k we have

$\frac{\delta(T_N-1)}{2} \leqslant \gamma - \bar{X}_{1T_N} \leqslant \frac{\delta(T_N+1)}{2}$ and for small δ

$(\gamma-\bar{X}_{1T_N})^2 \doteqdot (\delta T_N)^2/4$. Also

$$S_{1T_N} = \sum_{j=1}^{T_N} (x_j-\bar{X}_{1T_N})^2 = \delta^2 \sum_{j=1}^{T_N} \left\{\left(\frac{T_N+1}{2} - j\right)\right\}^2$$

$$= \frac{\delta^2 T_N(T_N^2-1)}{12}$$

It follows that

$$(\gamma-\bar{X}_{1T_N})^2/S_{1T_N} \doteqdot 3/\left(T_N-\frac{1}{T_N}\right) \doteqdot 3/T_N$$

and similarly it can be shown that $(\gamma-\bar{X}_{2T_N})^2/S_{2T_N} \doteqdot 3/(N-T_N)$, which simplifies 4.1(a) to

$$V(\hat{\gamma}^*_{T_N}) = \frac{4\sigma^2}{\beta^{*2}} \left(\frac{1}{T_N} + \frac{1}{N-T_N}\right),$$

which verifies Hinkley's result.

4.3 The asymptotic normality of $\hat{\gamma}^*_{T_N}$.

The following result (from Rao (13), p.387) for large sample theory is needed in order to establish the required result.

"Let $\hat{\underset{\sim}{\theta}}_N'$ be a k dimensional statistic $(\hat{\theta}_{1N},\ldots,\hat{\theta}_{kN})$ such that the asymptotic distribution of $\sqrt{N}(\hat{\theta}_{1N}-\theta_1),\ldots,\sqrt{N}(\hat{\theta}_{kN}-\theta_k)$ is k-variate normal with mean zero and variance matrix $\Sigma = (\sigma_{ij})$. Further let g be a function of k variables which is totally differentiable. Then the asymptotic distribution of $\sqrt{N}\,u_N = \sqrt{N}[g(\hat{\underset{\sim}{\theta}}_N)-g(\underset{\sim}{\theta})]$ is normal with mean zero and variance

$$V(\underset{\sim}{\theta}) = \sum_{i=1}^{k}\sum_{j=1}^{k} \sigma_{ij}\left(\frac{\partial g}{\partial \theta_i}\right)\left(\frac{\partial g}{\partial \theta_j}\right), \text{ provided } V(\underset{\sim}{\theta}) \neq 0.$$

If σ_{ij} and the partial derivatives of g are also continuous functions of $\underset{\sim}{\theta}$, then

$$\sqrt{N}\,u_N/\sqrt{V(\hat{\underset{\sim}{\theta}}_N)} \xrightarrow{L} X \sim N(0,1)."$$

For the sequence of configurations of the x values

being considered, we postulate that the following limits hold.

$$\lim_{N\to\infty} N\left(\frac{1}{T_N}+\frac{1}{N-T_N}+\frac{\bar{X}_1^2 T_N}{S_1 T_N}+\frac{\bar{X}_2^2 T_N}{S_2 T_N}\right) = k_{11}$$

4.3(a)
$$\lim_{N\to\infty} N\left(\frac{\bar{X}_1 T_N}{S_1 T_N}+\frac{\bar{X}_2 T_N}{S_2 T_N}\right) = k_{12}$$

$$\lim_{N\to\infty} N\left(\frac{1}{S_1 T_N}+\frac{1}{S_2 T_N}\right) = k_{22}$$

where k_{11}, k_{12}, k_{22} are some constants. It follows that

$$\lim_{N\to\infty} N\left(\frac{1}{T_N}+\frac{1}{N-T_N}+\frac{(\gamma-\bar{X}_1 T_N)^2}{S_1 T_N}+\frac{(\gamma-\bar{X}_2 T_N)^2}{S_2 T_N}\right)$$

$$= k_{11} - 2\gamma k_{12} + k_{22}\gamma^2 = k^* \quad \text{say.}$$

Then the variance matrix of $(\alpha^*_{T_N}, \beta^*_{T_N})$ approaches $\sigma^2\begin{pmatrix} k_{11} & k_{12} \\ k_{12} & k_{22}\end{pmatrix}$ as N increases. Putting $\hat{\theta}_{1N} = \hat{\alpha}^*_{T_N}$, $\hat{\theta}_{2N} = \hat{\beta}^*_{T_N}$ we have the stronger condition that $\sqrt{N}(\hat{\alpha}^*_{T_N}-\alpha^*)$, $\sqrt{N}(\hat{\beta}^*_{T_N}-\beta^*)$ have a bivariate normal distribution with mean zero and variance matrix

$$N\begin{bmatrix} \mathrm{Var}(\hat{\alpha}^*_{T_N}) & \mathrm{Cov}(\hat{\alpha}^*_{T_N},\hat{\beta}^*_{T_N}) \\ \mathrm{Cov}(\hat{\alpha}^*_{T_N},\hat{\beta}^*_{T_N}) & \mathrm{Var}(\hat{\beta}^*_{T_N}) \end{bmatrix}.$$

Let $g(\hat{\alpha}^*_{T_N},\hat{\beta}^*_{T_N}) = \hat{\alpha}^*_{T_N}/\hat{\beta}^*_{T_N} = \hat{\gamma}^*_{T_N}$ for $\hat{\beta}^*_{T_N} \neq 0$. Using the result from Rao, the asymptotic distribution of $\sqrt{N}(\hat{\gamma}^*_{T_N}-\gamma)$ is normal with mean zero and variance $N.V(\hat{\gamma}^*_{T_N})$. Also we have

$$\frac{\hat{\gamma}^*_{T_N}-\gamma}{\sqrt{V(\hat{\gamma}^*_{T_N})}} \xrightarrow{L} X \sim N(0,1).$$

One would hope as N increases that t_N approaches T_N so that the asymptotic properties of $\hat{\gamma}^*_{T_N}$ will also apply to $\hat{\gamma}^*_{t_N}$. We further hope that for the sequence of configurations of x values being considered we have $\hat{\gamma}^*_{t_N} = \hat{\gamma}_N$ so that the asymptotic properties of $\hat{\gamma}^*_{T_N}$ also apply to $\hat{\gamma}_N$.

For a fixed sample of size N and under the assumption that the above holds we proceed with the rest of this chapter.

4.4. Testing the Null Hypothesis $H_0 : \gamma = \gamma_0$

(a) From the asymptotic property it follows that under H_0 the test statistic $z = (\hat{\gamma}-\gamma_0)/\hat{\sigma}(\gamma_0)$ is approximately normally distributed with mean zero and variance one, where

$$\hat{\sigma}^2(\gamma_0) = \frac{\hat{\sigma}^2}{(\hat{\beta}_{2t_0}-\hat{\beta}_{1t_0})^2}\left[\frac{1}{t_0}+\frac{1}{N-t_0}+\frac{(\gamma_0-\bar{X}_{1t_0})^2}{S_{1t_0}}+\frac{(\gamma_0-\bar{X}_{2t_0})^2}{S_{2t_0}}\right], \hat{\beta}_{2t_0}\neq\hat{\beta}_{1t_0}$$

and $x_{t_0} \leqslant \gamma_0 < x_{t_0+1}$ with $\hat{\beta}_{1t_0}, \hat{\beta}_{2t_0}$ being the least squares estimates of β_1, β_2.

(b) The Likelihood Ratio Approach.

Firstly assuming σ^2 known, the likelihood ratio statistic is given by

$$\lambda = \frac{\exp[-(S_0-Z^2_{t_0}(\gamma_0))/2\sigma^2]}{\exp[-(S_0-Z^2_t(\hat{\gamma}))/2\sigma^2]}; \quad x_t \leqslant \hat{\gamma} < x_{t+1}$$

$\Rightarrow -2 \ln \lambda = (Z^2_t(\hat{\gamma})-Z^2_{t_0}(\gamma_0))/\sigma^2$ has an asymptotic χ^2 distribution with 1 degree of freedom. Substituting for σ^2 (by $\hat{\sigma}^2$) if it is unknown will not change the asymptotic result.

Without the assumption that σ^2 is known, we can write the likelihood ratio statistic as $\lambda = (\hat{\sigma}^2/\hat{\sigma}^2_0)$, where $\hat{\sigma}^2_0 = \hat{\sigma}^2(\gamma_0)$. It follows that

$$-2 \ln \lambda = N \ln(\hat{\sigma}^2_0/\hat{\sigma}^2)$$
$$= N \ln(1+(\hat{\sigma}^2_0-\hat{\sigma}^2)/\hat{\sigma}^2).$$

Since $\hat{\sigma}^2_0 \geqslant \hat{\sigma}^2$, with equality when $\hat{\gamma} = \gamma_0$, we have for γ_0 in the vicinity of $\hat{\gamma}$ that

$$-2 \ln \lambda \doteqdot N(\hat{\sigma}^2_0-\hat{\sigma}^2)/\hat{\sigma}^2$$

which has an asymptotic chi-squared distribution with 1 degree of freedom.

4.5 An Approximate Confidence Interval for γ.

(a) Using the asymptotic property, an approximate $100(1-\xi)$ per cent confidence interval is given by

$$\hat{\gamma} \pm z_{\xi/2} \cdot \sigma(\hat{\gamma}) \quad \text{where}$$

$$\frac{\xi}{2} = \int_{-\infty}^{-z_{\xi/2}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt \, .$$

(b) Another means of obtaining an approximate confidence interval is to consider the set of γ values which satisfy the equation $F(\gamma) = \frac{(\hat{\gamma}-\gamma)^2}{\hat{\sigma}^2(\gamma)} \leqslant \chi^2_\xi$, where χ^2_ξ is the upper $100\xi\%$ significant point of a chi-squared distribution with 1 degree of freedom. It is possible that such a confidence interval could be disjointed, depending on the function $F(\gamma)$. Graphically the confidence interval can be obtained by plotting $F(\gamma)$ over the values of γ and observing those ranges of γ for which $F(\gamma) \leqslant \chi^2_\xi$.

If the function $F(\gamma)$ is approximately parabolic in the neighbourhood of $\hat{\gamma}$, then an initial approximation to the approximate confidence interval is given by the roots of the quadratic $(\hat{\gamma}-\gamma)^2 = \chi^2_\xi . \hat{\sigma}^2(\gamma)$. This quadratic in γ can be expressed as $[\hat{\beta}-\chi^2_\xi.\text{Est Var}(\hat{\beta})]-2\gamma[\hat{\alpha}\hat{\beta}-\chi^2_\xi.\text{Est Cov}(\hat{\alpha},\hat{\beta})] + \gamma^2[\hat{\alpha}-\chi^2_\xi.\text{Est Var}(\hat{\alpha})] = 0$, where $\hat{\beta} = \hat{\beta}_2 - \hat{\beta}_1$ and $\hat{\alpha} = \hat{\alpha}_1 - \hat{\alpha}_2$.

CHAPTER 5

EXAMPLES

We analyse some experimental data sets and also use generated data sets to demonstrate various properties. The results for each data set will be set out under the following format.

(a) A discussion of the problem giving rise to the data, the data and a plot of the same.

(b) The procedures of Chapter 2 used to test for a split line model.

[1] Using differences between consecutive updated regression coefficients.

Two test results are given depending on the amalgamation procedure, the results being set out as follows.

'Analysis of Variance

Comparisons	∂f	Sums of Squares
Single slope	1	$Q_M \hat{\beta}_M^2$
Intersecting lines	1	$(M-2)\ \bar{d}^2$
Residual (d)	M-3	$\sum_{k=3}^{M} (d_k-\bar{d})^2$
Residual (y)	N-M	$\sum_{k=1}^{M} \sum_{k=1}^{m_i} (y_{ij}-\bar{y}_{i.})^2$
Total	N-1	$\sum_{i=1}^{M} \sum_{j=1}^{m_i} (y_{ij}-\bar{y}_{..})^2$

A test statistic, $t = (M-2)^{\frac{1}{2}}\ \bar{d}/s_d$, with s_d as discussed in §2.3, was calculated under the following scheme.

$$\text{if} \quad \begin{matrix} N = M \\ N - M \text{ small} \\ N - M \text{ large} \end{matrix} \quad \text{then} \quad t = \begin{matrix} t_1 \sim t_{N-3} \\ t_2 \sim t_{N-3} \\ t_3 \sim t_{N-M} \end{matrix}$$

If $N - M$ is large, we perform a further test by the use of

$$f = \frac{\sum_{k=3}^{M} (d_k - \bar{d})^2/(M-3)}{\sum_{i=1}^{M} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{i.})^2/(N-M)} \sim F_{M-3,N-M} .$$

[2] The Quadratic Component technique.

Analysis of Variance

Comparisons	∂f	Sum of Squares
Single slope	1	$Q_M \hat{\beta}_M^2$
Quadratic component	1	$R_Q - R_L$
Residual	N-3	$(N-3)S_Q$
Total	N-1	$\sum_{i=1}^{M} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}..)^2$

The test statistic, $f = \frac{(R_Q - R_L)}{S_Q} \sim F_{[1,N-3]}$ gives a test for non-linearity.

[3] The Graphic Approach which uses a moving average plot of the progressively updated intercept values.

[4] The Recursive Residual Approach

The plot of $[(W_k,k) : k = 3,...M]$ is represented on the graph by the following symbols.

A : W_k values

B : Upper Bound

C : Lower Bound

If the plot of W_k (A) crosses either boundary (B,C) then the Cusum technique indicates a non-linearity in the data.

(c) A plot of the residual function $S^2(u)$.

(d) The least squares estimates of α_1, α_2, β_1, β_2, γ, σ^2 and the times taken for each method of Chapter 3 are given. If the data plot shows clear non-linearity then methods 3.6, 3.7 are redone, using some set (ℓ_1,ℓ_2), to determine if any time saving can be achieved. The approximate 95 percent asymptotic confidence intervals for the intersection point (γ) completes the format. In data plots which strongly indicated non-linearity, section (b) on tests for non-linearity is omitted from the results for that data set.

5.1 EXAMPLE ONE: [Growth rate of human embryonic brain tissue; data kindly supplied by J. Dobbing.(5)]

Brief Glossary:

Neuron: An immature nerve cell.

Neuroblast: Any embryonic cell which develops into a nerve cell or neuron.

Neuroglia: (glia) The supporting structure of nervous tissue.

Myelin: A lipid substance forming a sheath around certain nerve fibres.

Myelination: The act of furnishing with or taking of myelin.

(a) From the time of conception until maturity the human body is subject to different growth spurts; the one we consider is the brain growth spurt. Poor nutrition during this period could inhibit normal brain development and consequently cause brain deficiencies.

A feature of mammalian brain growth is that neuroblast multiplication is almost complete before the major phase of glial multiplication begins, with the brain growth spurt beginning towards the end of the neuroblast stage. The first half of this spurt being glial, with the second half being mainly myelination.

Dobbing demonstrates (Example TWO) that fast and slow growing animals undergo the process of cell division, myelination and growth in brain weight at fixed chronological ages. Since this result can also be applied to humans it is of some importance to determine the point in time at which the brain growth spurt begins.

Dobbing demonstrated that this occurs at about the eighteenth week of gestation. The study was based on measurements from 148 human brains recovered from aborted foetuses, accident victims and children who died from short illnesses. No brain was accepted which showed any conspicuous neuropathology.

Each brain was divided into three regions (Forebrain, Cerebellum and Stem) and measurements of the DNA-P, water and cholesterol content were made. A plot of log e (DNA-P) from forebrain against age for brains from 10 gestational weeks to 4 postnatal months shows a clear non-linear trend. Within this age limit there were N=106 data value, with M=42.

Data

Age (Weeks)	log e DNA in forebrain	Age (Weeks)	log e DNA in forebrain
10	2.045	38	6.568,6.299,6.705,
11	2.269		6.535,6.679,6.446,
12	3.120,1.721		6.547
13	3.157	39	6.967
14	4.431	40	6.388,6.438,6.425,
15	4.071		6.719,6.425
16	4.290,4.868,	41	6.518,6.524,6.413,
	4.511,4.875		6.521
17	5.017,5.165,	42	6.346
	5.313,4.691	43	6.592,6.486
18	5.204,5.081,5.118,	44	6.468
	5.517,5.407	46	6.442,6.507,6.765,
19	5.561,5.591,5.421		6.823,6.399,6.588,
20	5.583,5.613		6.416
21	5,908,5.796	47	6.740,6.826
22	5.606	48	7.064,6.690,7.154,
25	5.971		6.777,6.802
26	5.793,5.883	49	7.098,6.516
27	5.883,5.841,6.165	50	7.034,6.746
28	6.076,5.897,5.930	51	6.570
29	5.971	52	6.642
30	6.188,6.144,6.009,	53	6.999,6.630,6.416,
	6.211,6.242		7.183,6.941,6.405,
31	6.131		6.927,6.436
32	6.396,6.117,6.013	54	6.971
33	6.625,6.588,6.161	55	6.689
34	6.203	57	6.819,6.813,7.134
35	6.240		

A Data Plot of $\log_e$(DNA-P) for Forebrain versus Age.

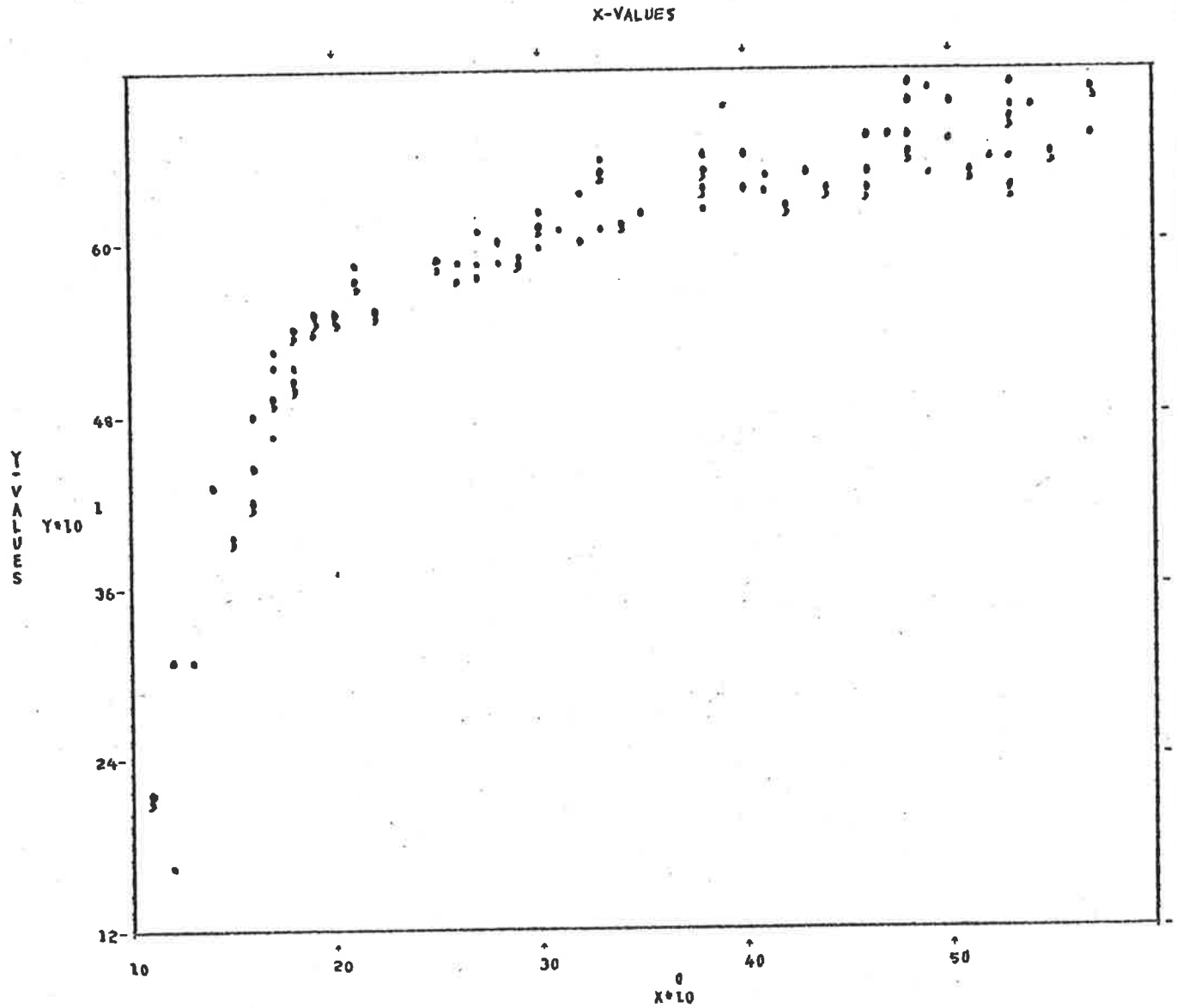


Fig. 5.1.(1)

(c) A plot of the overall residual sum of squares function, $S^2(u)$.

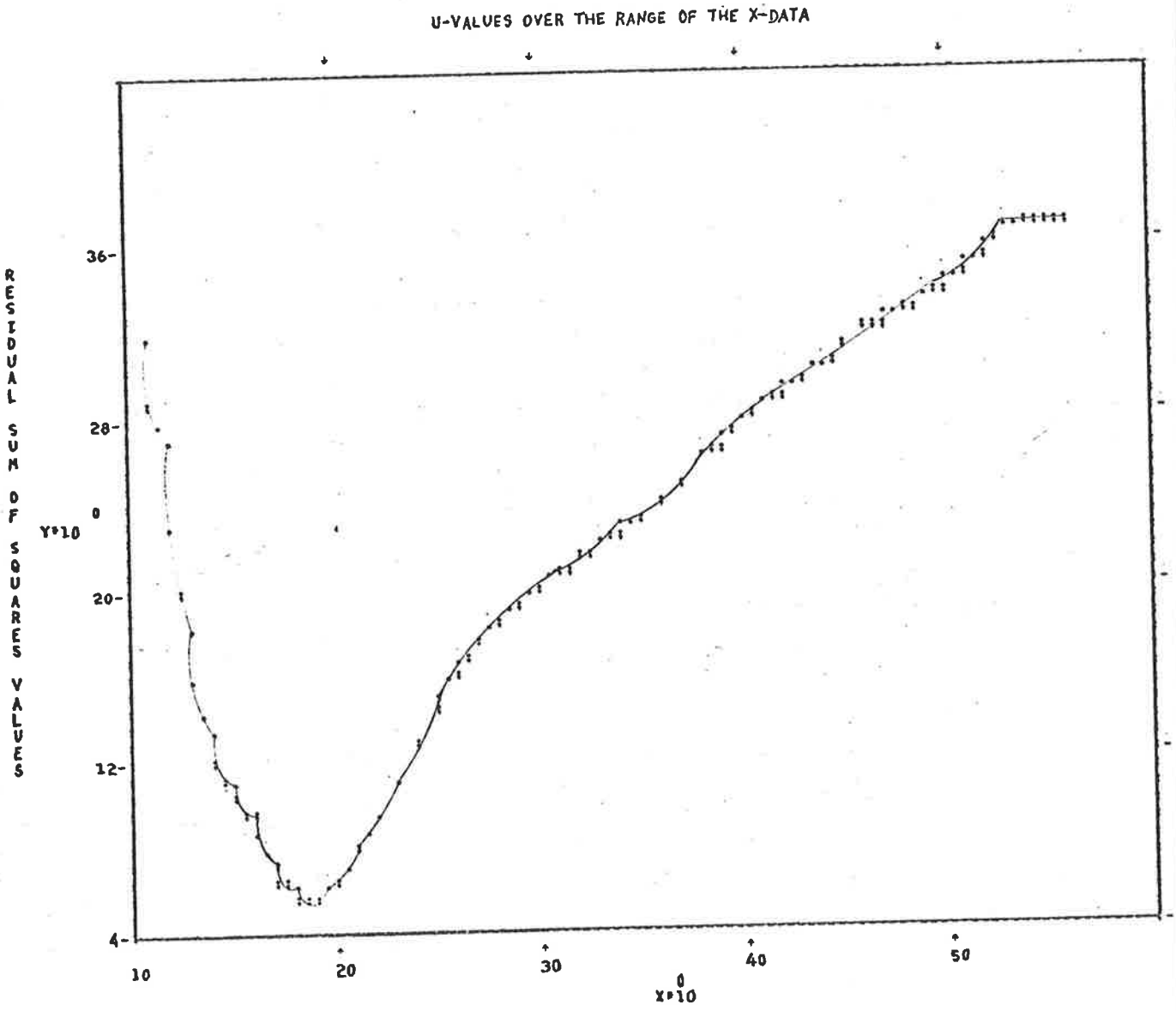


Fig. 5.1.(2)

(d)

Parameter	Estimate	Standard Error
α_1	-2.47	.34
α_2	5.09	.10
β_1	4.38×10^{-1}	$.22 \times 10^{-1}$
β_2	3.40×10^{-2}	$.25 \times 10^{-2}$
γ	18.71	.27

and $\hat{\sigma}^2 = 6.04 \times 10^{-1}$

Method		Time (sec $\times 10^{-2}$)
Hudson	§3.4	5.5
Hinkley	§3.5	5.2
Upper bound	§3.6	5.1
Interval	§3.7	5.0

Since non linear trend was evident in the data we chose $\ell_1 = 7$, $\ell_2 = 16$ which gives an interval (16,27). Using this interval as an initial interval for methods §3.6 and §3.7 we obtained the following:

Method		Time (sec $\times 10^{-2}$)
Upper bound	§3.6	4.9
Interval	§3.7	4.5

An approximate 95 per cent asymptotic confidence interval is given by (18.19, 19.23).

This verifies Dobbing's result that the brain growth spurt begins at about the eighteenth week of gestation with glial multiplication carrying on well into the second postnatal year. Since under-nourishment during pregnancy will not begin to seriously retard foetal growth till the third trimester (Dobbing) the neuronal multiplication phase in humans would be spared. Such later restriction will equally certainly not spare the establishment of the network of synoptic connectivity, and if such restriction is continued through the first two postnatal years, then poor brain development could be expected.

5.2 EXAMPLE TWO: [Environmental effects on the onset of the brain growth spurt; data kindly supplied by J. Dobbing.]

To demonstrate that the timing of the mammalian brain growth spurt is unaffected by maternal malnutrition, measurements on two groups of rats were made. One group consisted of rats born into large litters, the other group of rats born into small litters. A data plot of the $\log_e$DNA-P content for whole brain against age (0-196 days) shows a clear non-linearity for both groups (Fig. 5.2.(1); 5.2.(2)).

(a) Small litter group consisted of N=109 rats, with M=26.

Data

Age (days)	$\log_e$ DNA-P whole brain	Age (days)	$\log_e$ DNA-P whole brain
0	.349	32	1.822,1.694,1.795,
2	.294, .352		1.678,1.585,1.557,
4	.811,1.019		1.569,1.613
6	1.084,1.095,1.0675	34	1.543,1.611,1.610
8	.982,1.123, .956	35	1.611,1.623,1.547
10	1.413,1.403	36	1.709,1.483,1.483
12	1.506,1.575	38	1.632,1.591,1.639
14	1.673,1.655	42	1.556,1.575,1.506,
16	1.819,1.671,1.696,		1.839,1.623,1.578,
	1.654,1.526		1.501,1.715,1.535
18	1.616,1.655,1.597,	49	1.617,1.637,1.571,
	1.854		1.595,1.576,1.555,
20	1.886,1.800,1.741,		1.550,1.511
	1.749,1.835,1.777	70	1.731,1.683,1.868,
22	1.757,1.579		1.925,1.705,1.573,
24	1.632,1.652,1.812		1.679
26	1.703,1.687,1.674	196	1.963,1.899,1.841,
28	1.605,1.793,1.579,		1.679,1.571,1.725,
	1.813		1.942,1.787,1.668,
30	1.765,1.856,1.691,		1.681,1.679,1.780
	1.612,1.750		1.863,1.607,1.649,
			1.627

Data plot of log e DNA-P for whole brain versus age (0-196 days) for rats born into small litters.

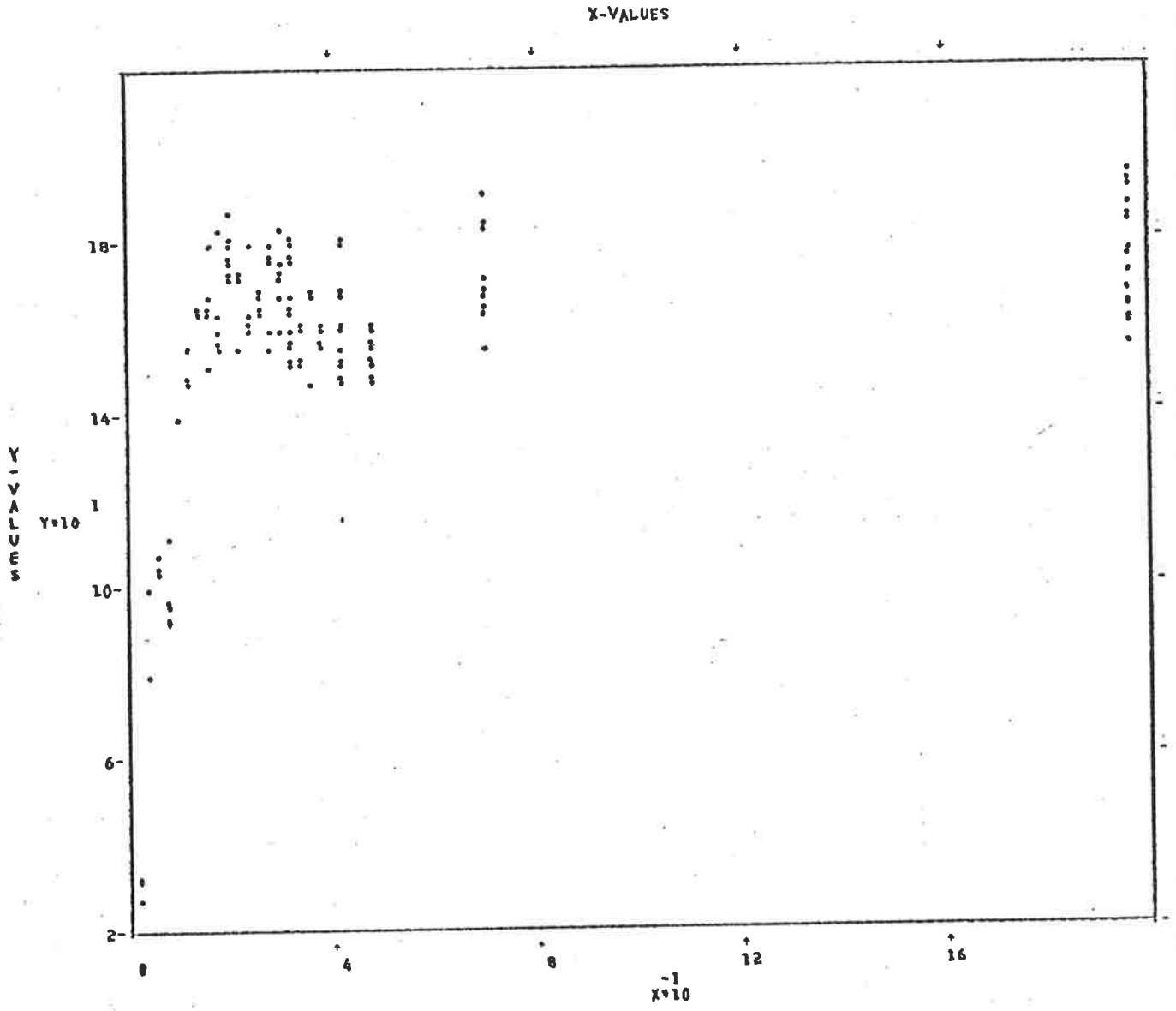
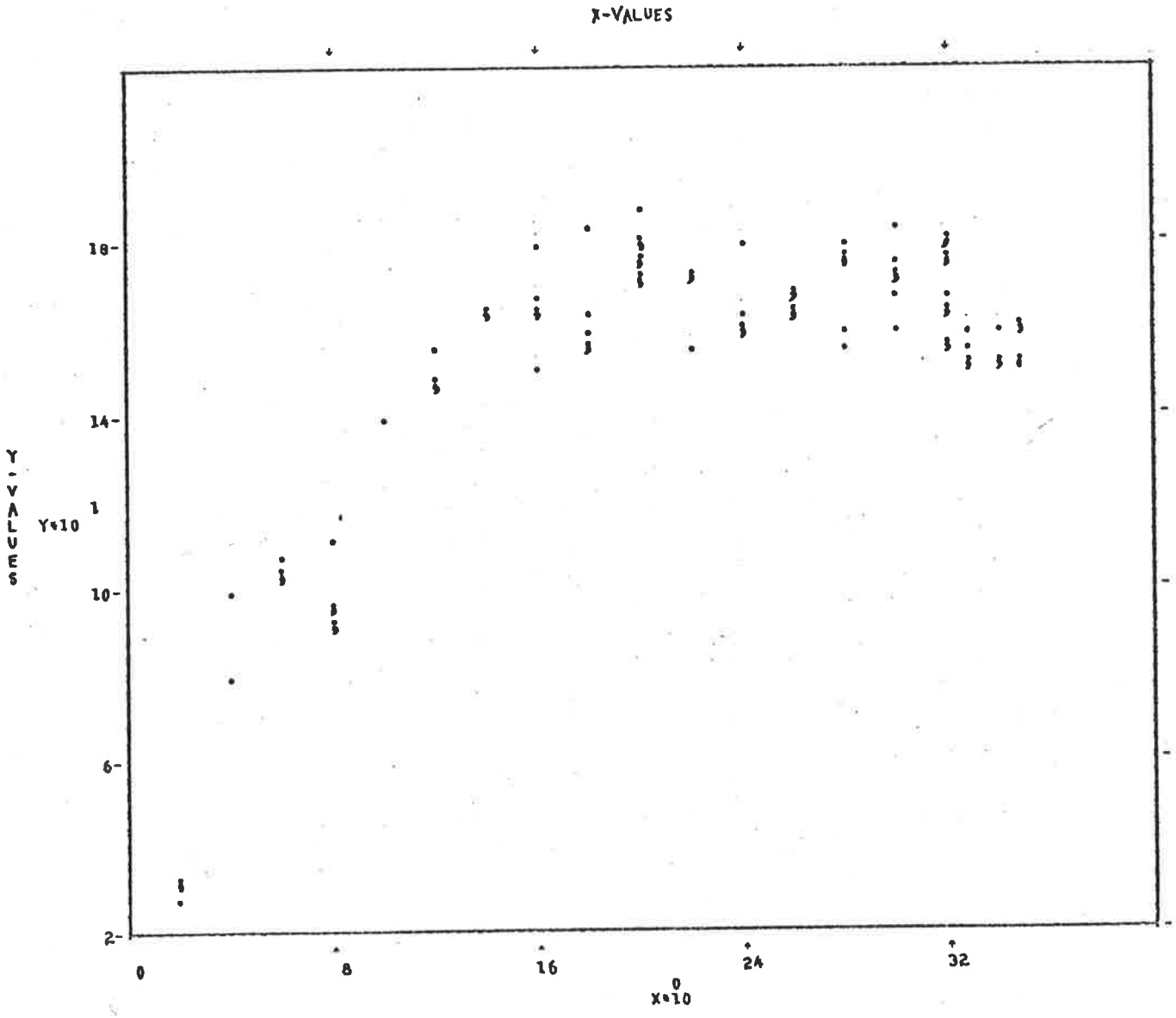


Fig. 5.2.(1)

A clearer picture as to the applicability of a split-line model to this data can be seen from a plot of the data over the age limits, 0-35 days.

Data plot of log e DNA-P for whole brain versus age (0-35 days) for rats born into small litters.



The large litter group consisted of 157 animals, with M=25.

Age (days)	log e DNA-P whole brain	Age (days)	log e DNA-P whole brain
0	.349	30	1.314,1.472,1.482, 1.446,1.494,1.505, 1.409,1.501,1.442, 1,362,1.369,1.584
2	.294, .352		
4	.793, .676, .950, .951, .997, .933	32	1.488
6	.873, .851,1.011	33	1.349,1.456,1.501, 1.649,1.593,1.501
8	1.174, .962,1.096, 1.130,1.044, .980	34	1.543,1.6, 1.576, 1.531,1.495
11	1.313,1.121,1.158, 1.106,1.282	36	1.541,1.423,1.302, 1.545,1.499,1.459
12	1.404,1.491,1.523, 1.395,1.347,1.408	37	1.516,1.522,1.525, 1.443
14	1.391,1.265,1.319, 1.233,1.428,1.148	42	1.292,1.398,1.628, 1.427,1.456,1.375, 1.507,1.556
16	1.373,1.325,1.640, 1,200,1.329,1.335, 1.291,1.227	49	1.830,1.491,1.669, 1.444,1.412
18	1.483,1.556,1.507, 1.628,1.448	70	1.705,1.562,1.692, 1.561,1.394,1.551, 1.659
20	1.491,1.293,1.550, 1.552,1.493,1.502, 1.185,1.517	196	1.635,1.459,1.520, 1.504,1.708,1.413, 1.509,1.668,1.502, 1.571,1.440,1.587, 1.533,1.773,1.723, 1.797,1.579,1.535, 1.672,1.601,1.479
22	1.555,1.678,1.492, 1.471		
24	1.511,1.568,1.520, 1.521,1.476,1.547, 1.456,1.507,1.492		
26	1.319,1.327,1.374, 1.519,1.551,1.359, 1.288,1.427,1.686		
28	1.647,1.683,1.668, 1.458		

Data plot of log e DNA-P for whole brain versus age (0-196 days) for rats born into large litters.

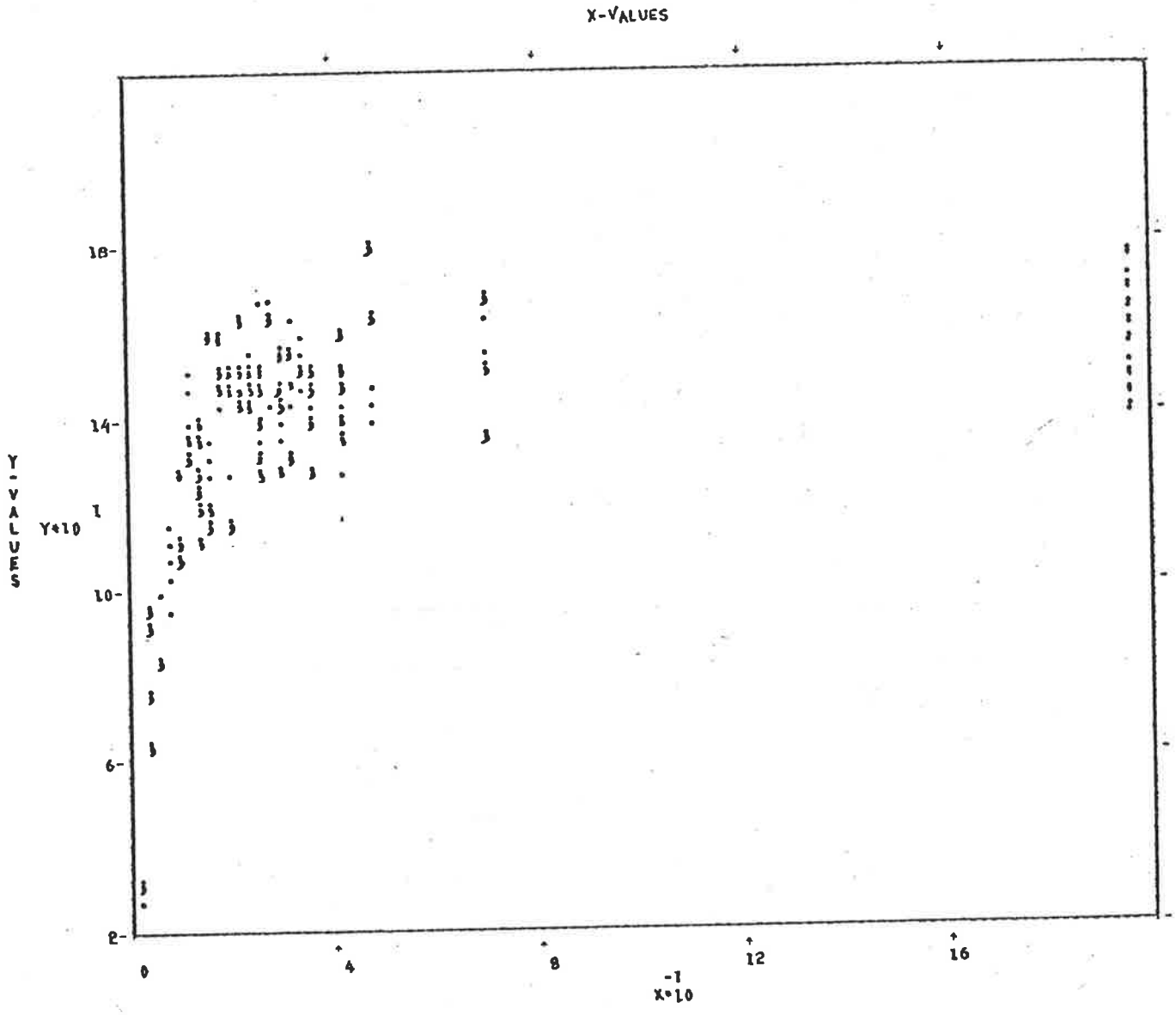
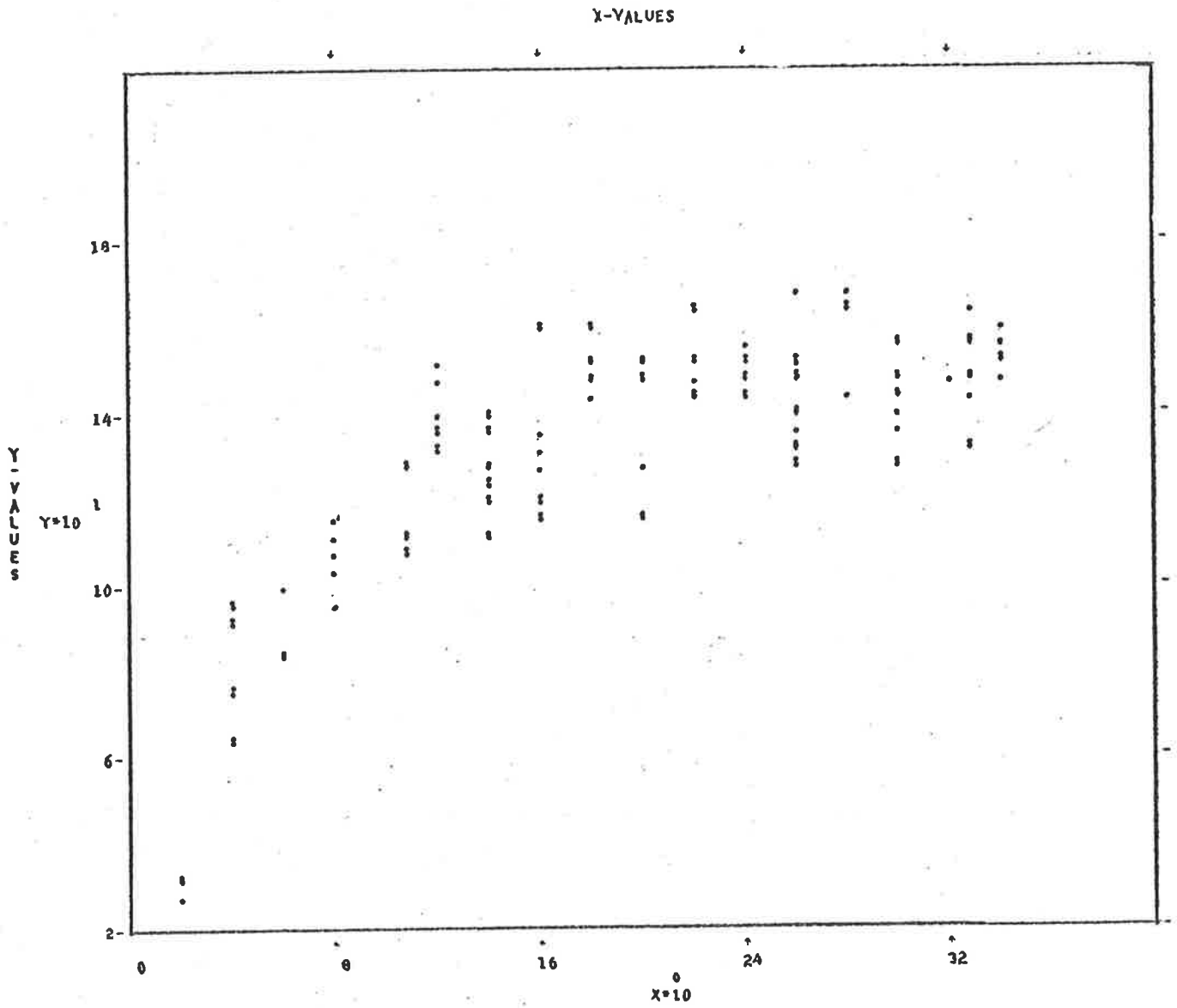


Fig. 5.2.(2)

As in Fig. 5.2.(1), a clearer picture as to the applicability of a split-line model can be seen from a plot over the age limits, 0-34 days.

Data plot of $\log_e$ DNA-P for whole brain versus age (0-34 days) for rats born into large litters



(c) A plot of the overall residual sum of squares function for the data from the small litters group.

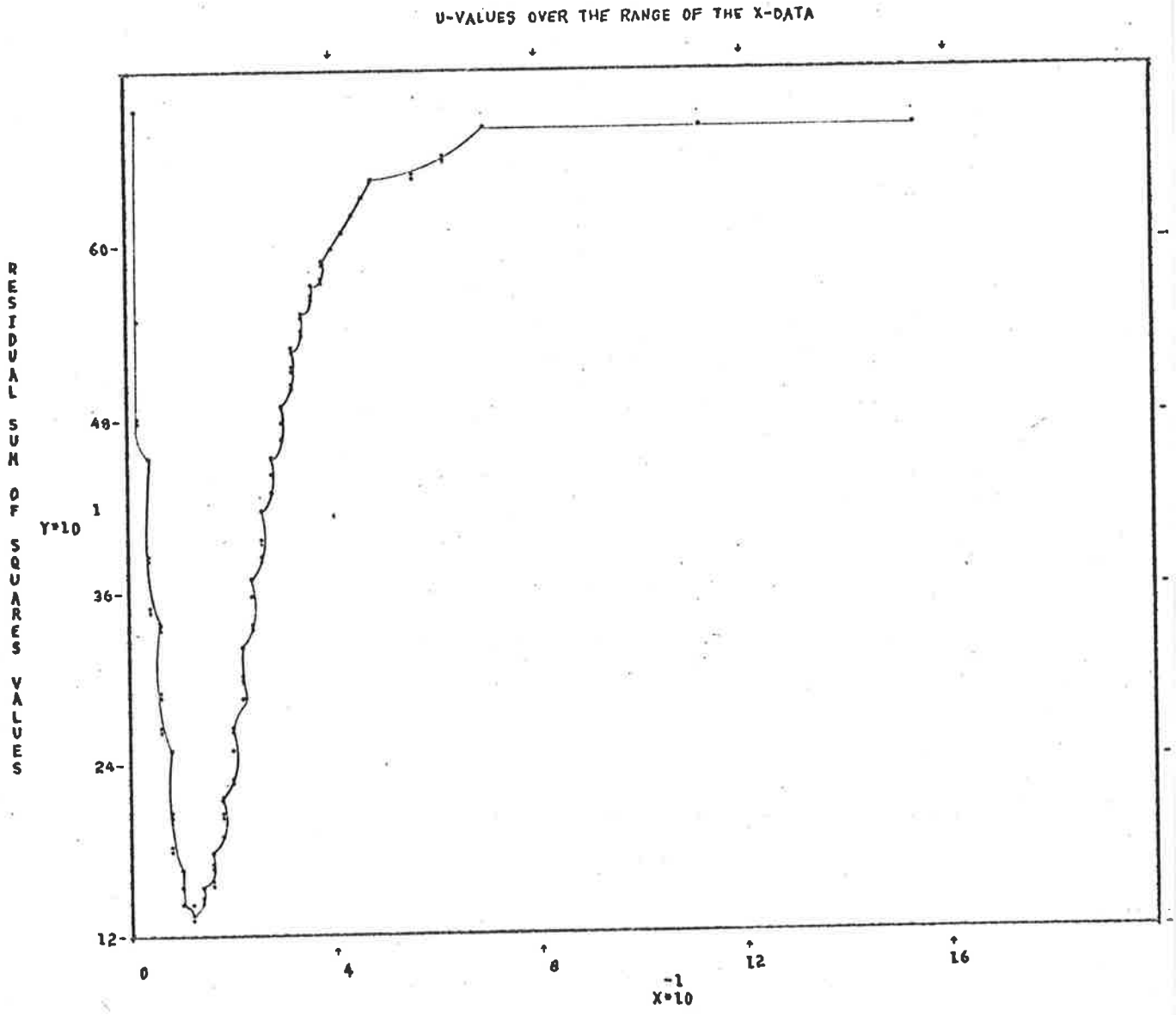


Fig. 5.2.(3)

A plot of the overall residual sum of squares function for the data from the large litters group.

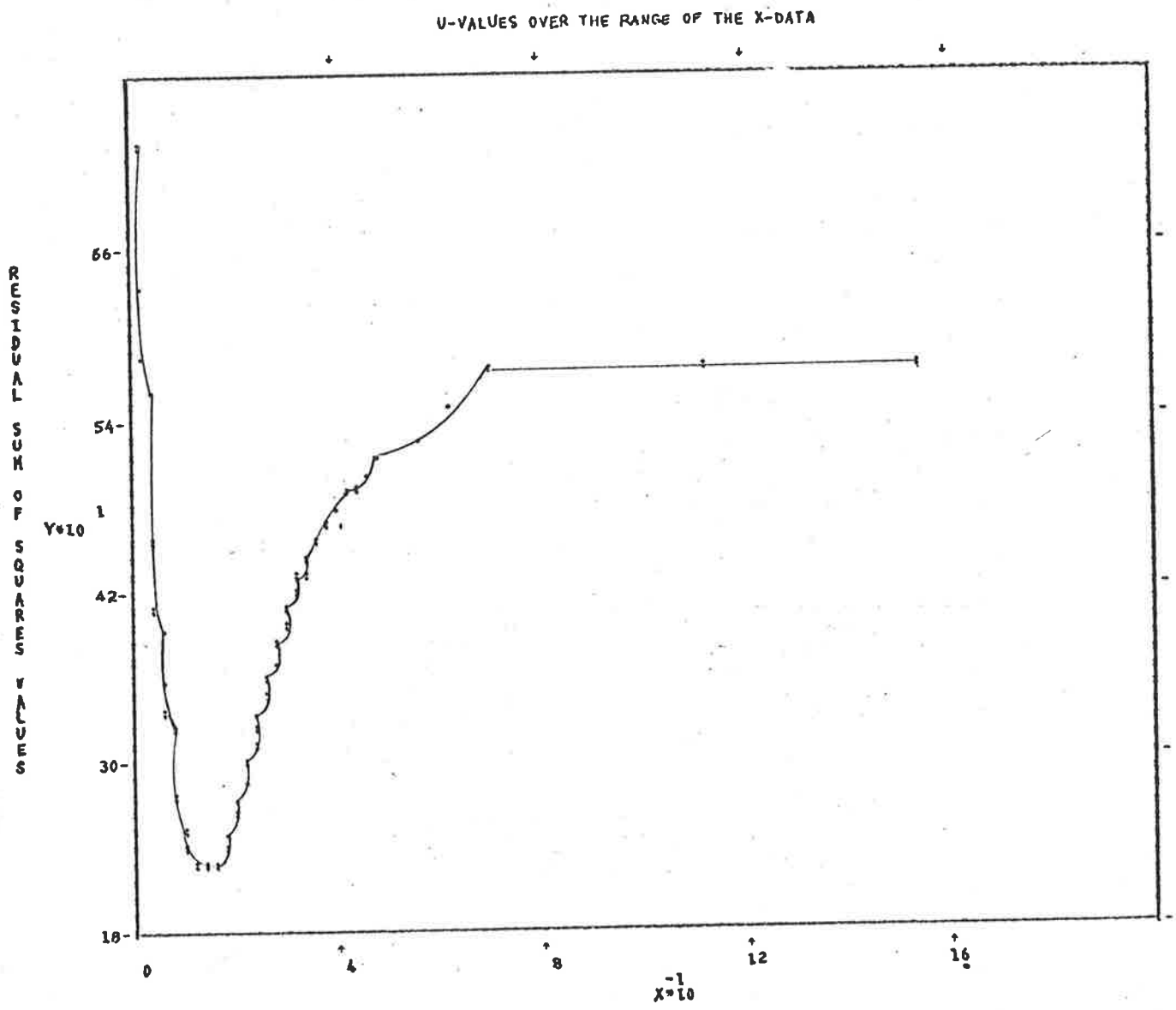


Fig. 5.2.(4)

(d) Small litters group

Parameter	Estimate	Standard Error
α_1	3.23×10^{-1}	$.63 \times 10^{-1}$
α_2	1.65	$.17 \times 10^{-2}$
β_1	1.04×10^{-4}	$.85 \times 10^{-3}$
β_2	4.54×10^{-1}	$.19 \times 10^{-1}$
γ	12.82	.61 ×

and $\hat{\sigma}^2 = 1.36 \times 10^{-2}$

Method		Time (sec × 10^{-2})
Hudson	§3.4	4.5
Hinkley	§3.5	4.1
Upper bound	§3.6	4.3
Interval	§3.7	4.5

Using $\ell_1 = 6$, $\ell_2 = 9$ to obtain an initial interval (10,16), we obtain

Method		Time (sec × 10^{-2})
Upper bound	§3.6	4.2
Interval	§3.7	4.0

An approximate 95 percent asymptotic confidence interval is given by (11.62, 14.02).

Large litter group

Parameter	Estimate	Standard Error
α_1	4.28×10^{-1}	$.53 \times 10^{-1}$
α_2	1.45 ×	$.15 \times 10^{-1}$
β_1	7.94×10^{-2}	$.63 \times 10^{-2}$
β_2	7.55×10^{-4}	$.17 \times 10^{-3}$
γ	12.98	.55

and $\hat{\sigma}^2 = 1.51 \times 10^{-2}$

Method		Time (sec × 10^{-2})
Hudson	§3.4	4.9
Hinkley	§3.5	4.8
Upper bound	§3.6	4.6
Interval	§3.7	4.6

Using $\ell_1 = 5$, $\ell_2 = 10$ to obtain an initial interval (8, 18), we obtain

Method		Time (sec × 10^{-2})
Upper bound	§3.6	3.7
Interval	§3.7	3.8

An approximate 95 percent asymptotic confidence interval is given by (11.91,14.05).

Conclusion

An approximate asymptotic test for the null hypothesis that the intersection points for both groups are equal is given by

$$z = \frac{12.98 - 12.82}{\sqrt{0.55^2 + 0.61^2}} = \frac{0.16}{\sqrt{.68}} = 0.2,$$

where z is a standard normal deviate.

It follows that the experiment does not contradict Dobbings initial contention, namely that the timing of the brain growth spurt in rats is independent of environmental conditions.

Therefore since no catch-up facility exists in the brain once the brain growth has ceased, the results of environmental factors which inhibit brain growth during the spurt period could be a deficiency of mental ability.

5.3 EXAMPLE THREE: [A light sensitivity experiment, data supplied by H. Wainer.]

(a) In an experiment to determine a subject's sensitivity to the brightness of a stimulus light, the subject was initially placed in a dark room for 15 minutes. The subject then stared directly at a 300 W. bulb for 5 minutes from a distance of 25 cm. After this preadaptation period the stimulus light, the brightness of which could be regulated in several ways, was introduced and the perceived intensity recorded. The method is given in Wainer (14).

Here we have N=M=30 and

x : Time, in minutes, from the end of preadaptation time.

y : - log (subjective intensity).

Data

x	y	x	y	x	y
.5	1.52	6.5	3.30	13.9	3.95
1.0	1.85	7.5	3.50	14.5	4.05
1.5	2.00	8.5	3.37	15.2	4.10
2.0	2.20	9.1	3.70	15.6	4.10
2.25	2.50	9.8	3.72	16.5	4.10
3.0	2.40	10.5	3.72	17.0	4.31
3.5	2.50	11.4	3.74	17.6	4.45
3.8	2.80	12.0	3.76	18.8	4.43
4.1	3.20	12.5	3.80	19.4	4.50
5.9	3.10	12.7	4.00	20.8	4.50

A plot of -log (subjective intensity) versus time (min.) since the end of the preadaptation period.

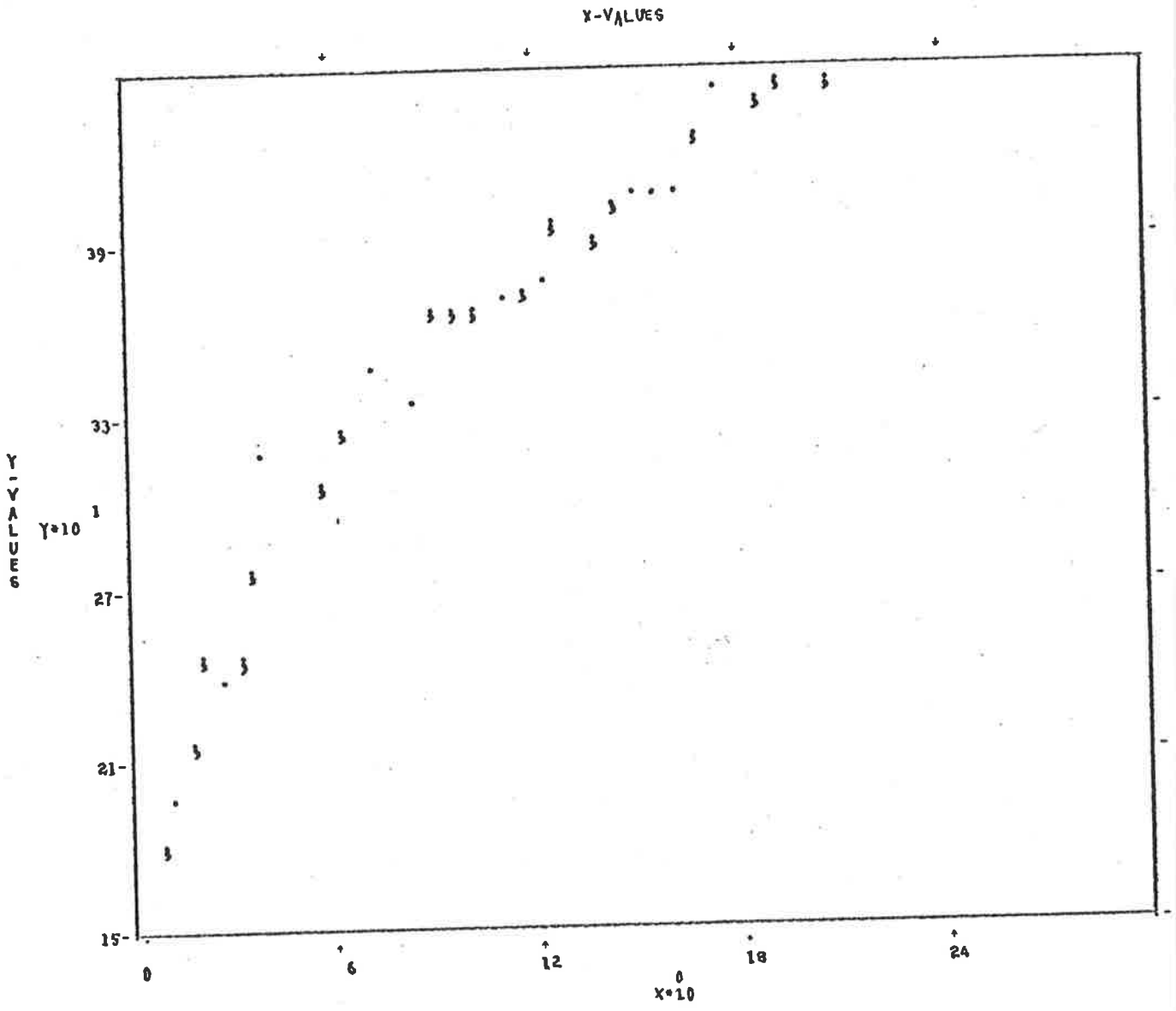


Fig. 5.3.(1)

(c) A plot of the overall residual sum of squares function, $S^2(\mu)$.

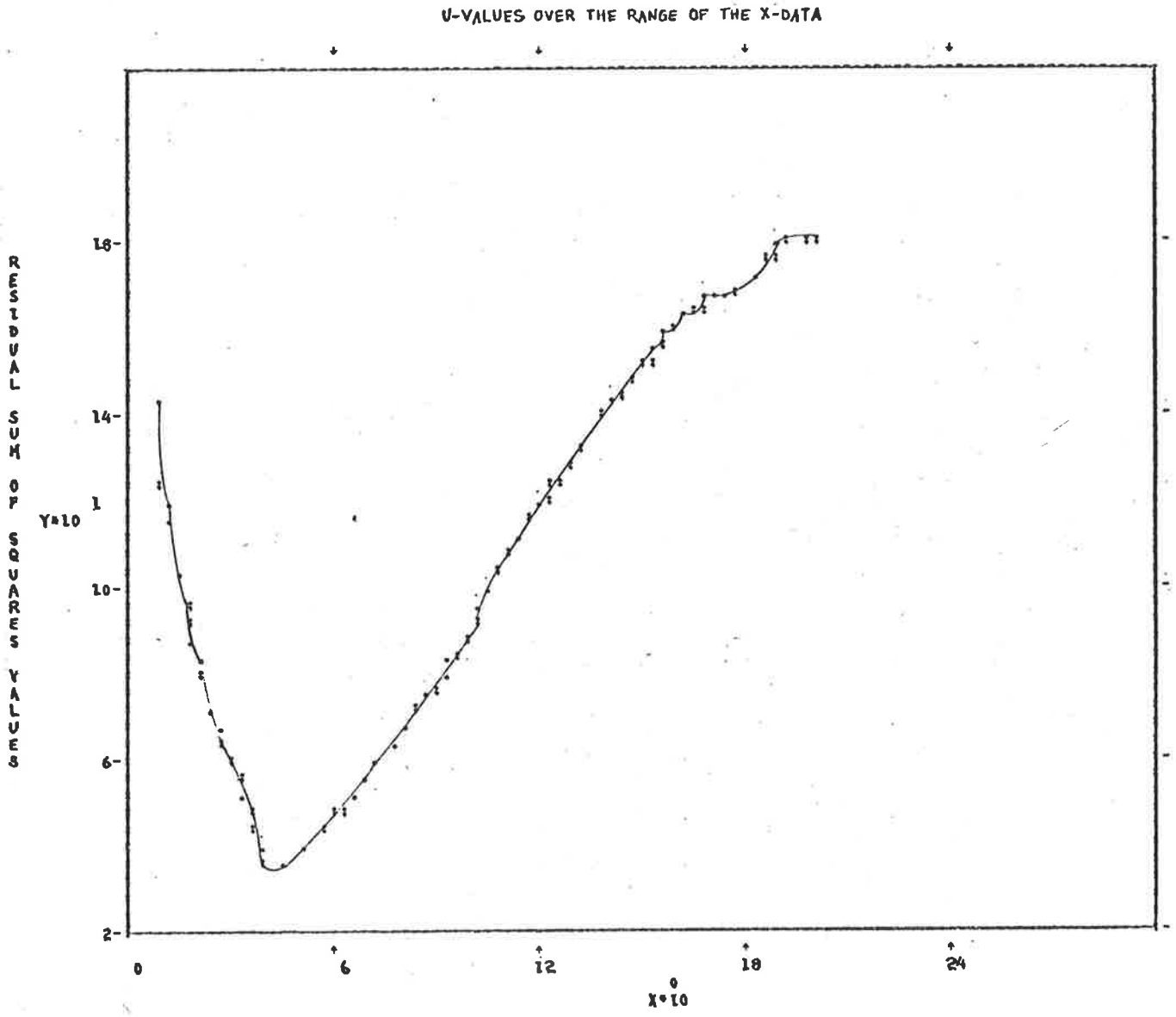


Fig. 5.3.(2)

(d)

Parameter	Estimate	Standard Error
α_1	1.42	$.88 \times 10^{-1}$
α_2	2.73	$.84 \times 10^{-1}$
β_1	3.78×10^{-1}	$.32 \times 10^{-1}$
β_2	9.01×10^{-2}	$.61 \times 10^{-2}$
γ	4.56	.35

and $\hat{\sigma}^2 = 1.38 \times 10^{-2}$

Method		Time (sec $\times 10^{-2}$)
Hudson	§3.4	4.4
Hinkley	§3.5	4.1
Upper bound	§3.6	4.0
Interval	§3.7	3.9

Using $\ell_1 = 6$, $\ell_2 = 12$ which gives an initial interval of (3,7.5) we obtain:

Method		Time (sec $\times 10^{-2}$)
Upper bound	§3.6	3.9
Interval	§3.7	3.8

An approximate 95 percent asymptotic confidence interval is given by (3.88,5.23). For an interpretation of these results, see Wainer (14).

5.4 EXAMPLE FOUR: [Stagnant surface layer height in water flows; data supplied by D.W. Bacon and D.G. Watts.]

In an investigation of the behaviour of stagnant surface layer height in a controlled flow of water down an inclined channel for a particular surfactant the following data was obtained, where

x : log (flow rate in gall/cm sec.)

y : log (band height in cm)

A plot of the data demonstrates the applicability of a split-line model. In this experiment 48 measurements were made, with M = 20.

x	y	x	y	x	y
-1.51	1.15	-.40	.69	.21	.34
	1.13		.55	.29	.31
-1.39	1.10		.60		.31
	1.06		.61	.33	.17
-1.08	.93	-.25	.58	.44	.14
	.95		.55	.46	.10
	.99	-.12	.52		.10
	.99		.54	.64	-.12
- .8	.87	0.01	.44	.80	-.33
	.89		.49	.85	-.36
	.89		.44		-.36
	.83		.46	1.03	-.60
	.80	.11	.34		-.56
- .58	.67		.42	1.19	-.73
	.73		.40		-.73
- .40	.64		.39		-.80

A plot of the log (flow rate) versus log (band height).

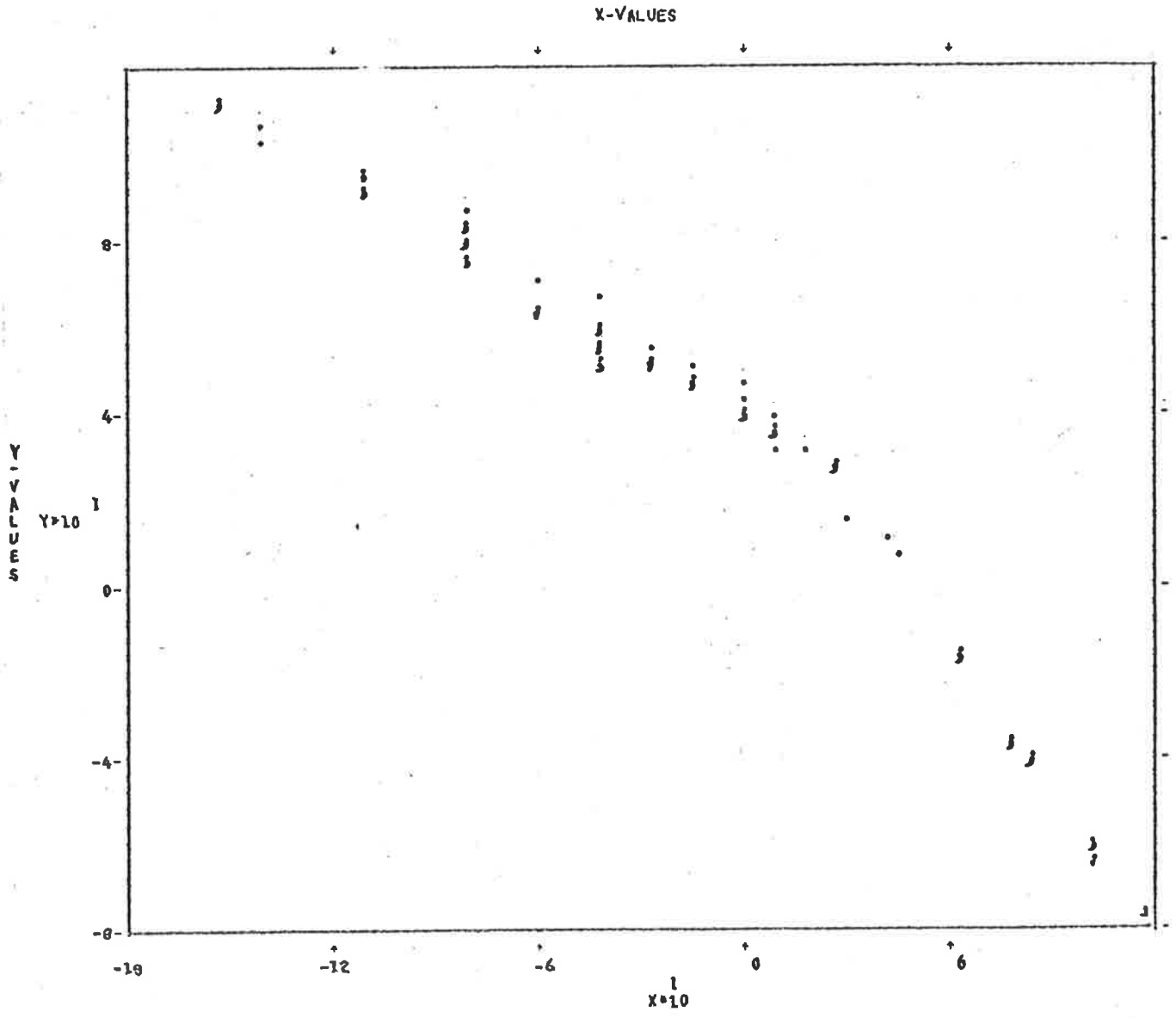


Fig. 5.4.(1)

(c) A plot of the overall residual sum of squares function, $S^2(\mu)$.

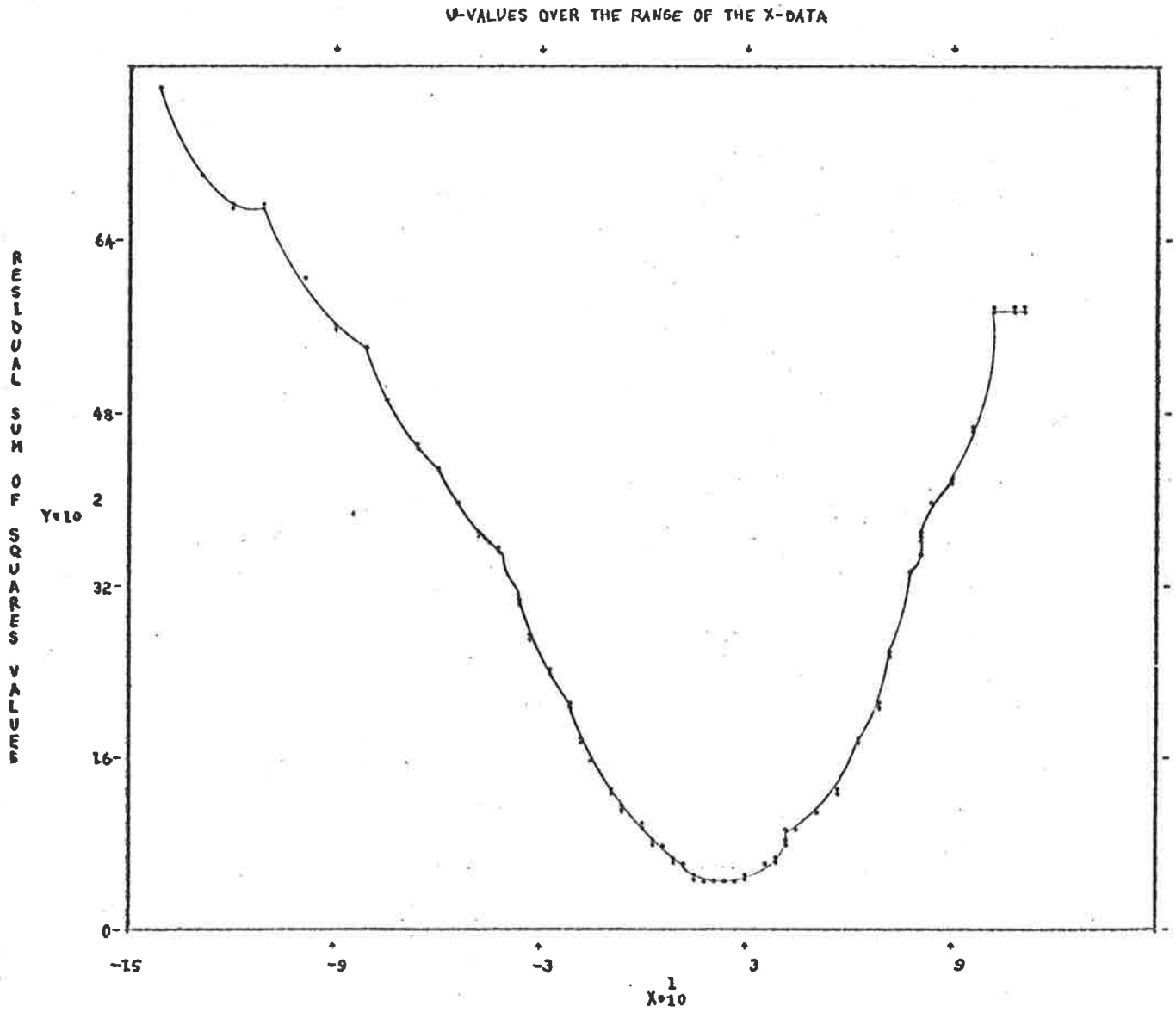


Fig. 5.4.(2)

(d)

Parameter	Estimate	Standard Error
α_1	.45	$.85 \times 10^{-2}$
α_2	.63	$.21 \times 10^{-1}$
β_1	- .47	$.12 \times 10^{-1}$
β_2	-1.16	$.27 \times 10^{-1}$
γ	.25	$.27 \times 10^{-1}$

and $\hat{\sigma}^2 = 1.18 \times 10^{-3}$

Method		Time (sec $\times 10^{-2}$)
Hudson	§3.4	4.0
Hinkley	§3.5	3.6
Upper bound	§3.6	3.5
Interval	§3.7	3.8

Using $\ell_1 = 5$ and $\ell_2 = 15$, which gives an initial interval (-.58,.44) we obtain:

Method		Time (sec $\times 10^{-2}$)
Upper bound	§3.6	3.5
Interval	§3.7	4.0

An approximate 95 percent asymptotic confidence interval is given by (.2,.31). For an interpretation of these results, see Bacon and Watts (1).

5.5 EXAMPLE FIVE

(a) For equally spaced x values over the range from 0 to 22, 100 data points were generated for the following split-line model.

$$E(y) = \begin{cases} 2.6 + 0.7x : & x \leqslant 12 \\ 5 + 0.5x : & x \geqslant 12 \end{cases}$$

where $\sigma^2 = 1$, $\gamma = 12$ and from the data we have that $x_{55} = \gamma = 12$.

This data is used to demonstrate the following points:

1. From the plot no visual evidence suggests non-linearity. This data, then, will try out the effectiveness of the tests in Chapter 2.
2. Applying Hinkley's criteria (3.3.2) we see that the estimation procedure could be ill defined since

$$\eta_T|B| = \left|\frac{\beta_2 - \beta_1}{\sigma}\right| \eta_T = 4.95 < 5. \qquad \text{(c.f. Ch. 3)}$$

Since we consider equally spaced x values, we have $N = M = 100$.

Plot of the generated data from the model

$$E(y) = \begin{cases} 2.6 + 0.7x & : x \leqslant 12 \\ 5 + 0.5x & : x \geqslant 12 \end{cases}$$

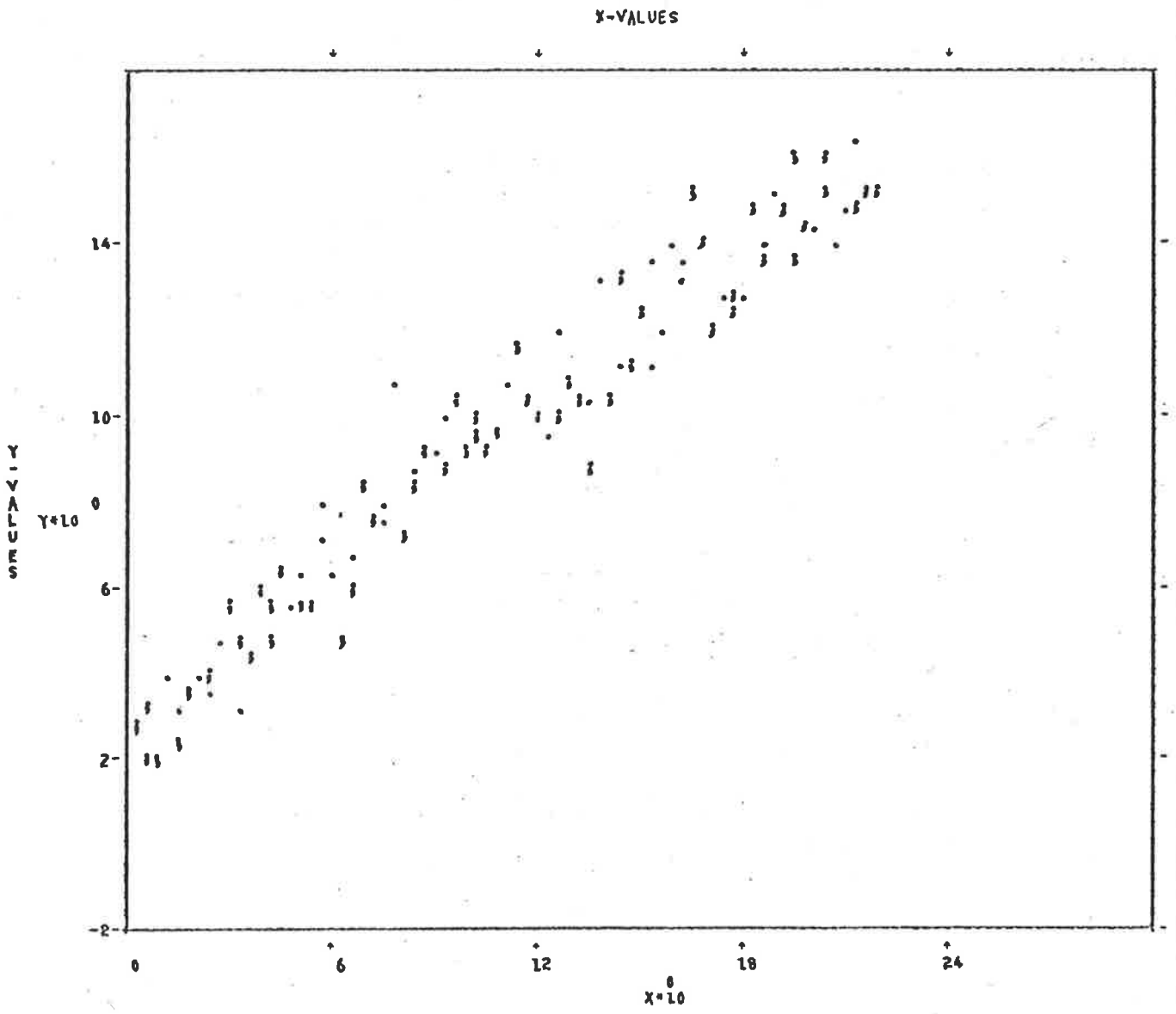


Fig. 5.5.(1)

(b) [1] Using differences between consecutive up-dated regression coefficients.

(i) Left to Right amalgamation

A.O.V.

Comparison	∂f	SS	MSS
Single slope	1	1547.75	1547.75
Intersecting lines	1	20.48	20.48
Residual (y)	97	87.01	0.9
Total	99	1655.24	

With $(M-2)^{\frac{1}{2}}\bar{d} = 4.53$, the test statistic

$$t_1 = \frac{4.53}{\sqrt{0.9}} = 4.78$$

gives a significant result for testing the non-linearity of the data.

(ii) Right to Left amalgamation

A.O.V.

Comparison	∂f	SS	MSS
Single slope	1	1547.75	1547.75
Intersecting lines	1	8.42	8.42
Residual (y)	97	99.07	1.02
Total	99	1655.24	

and $t_1 = \frac{2.9}{\sqrt{1.02}} = 2.87 \sim t_{97}$ gives a significant result for non-linearity of the data.

[2] The Quadratic Component technique

A.O.V

Comparison	∂f	SS	MSS
Single slope	1	1547.75	1547.75
Quadratic component	1	18.61	18.61
Residual	97	88.88	0.92
Total	99	1655.24	

The test statistic $f = \frac{18.61}{0.92} \sim F_{1,97}$ gives a significant result for non-linearity in the data.

[3] The graphic approach

A plot of the moving average of updated intercept values when amalgamating from left to right.

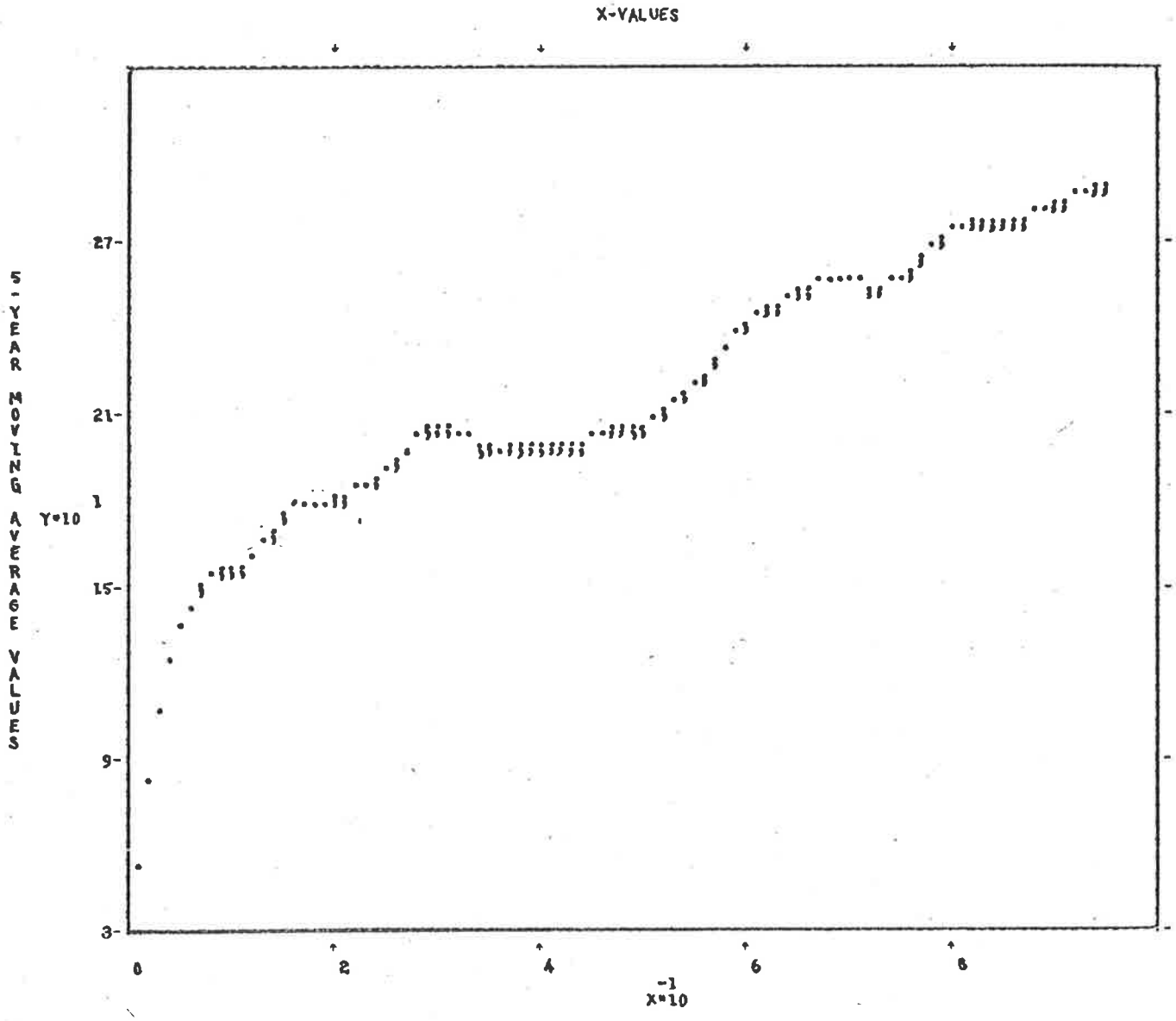


Fig. 5.5.(2)

The persistent upward movement in the moving average plot is very indicative of non-linearity in the data.

[4] The Recursive Residual Approach, using amalgamation from left to right.

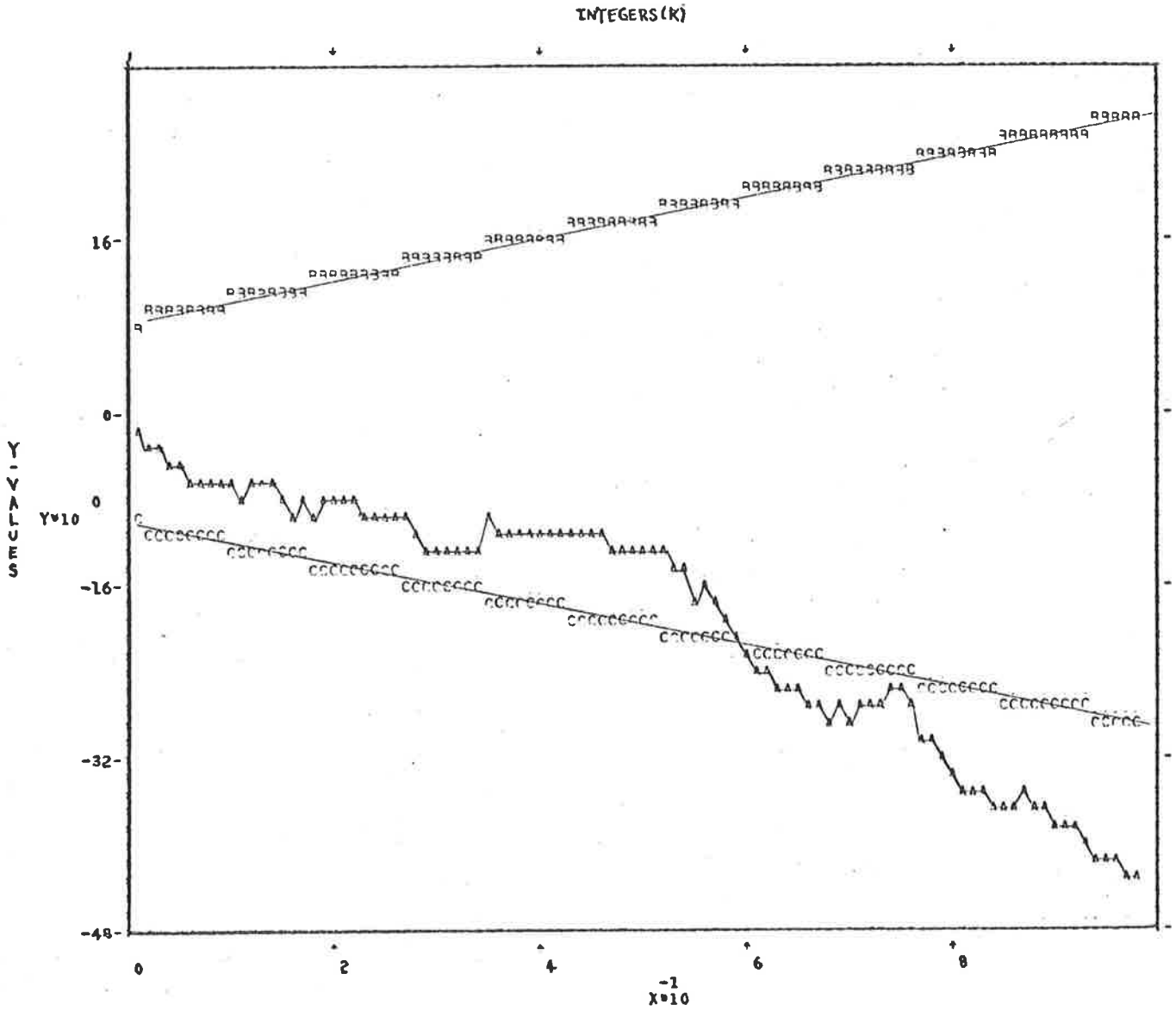


Fig. 5.5.(3)

From above, the plot (points: A) crosses the lower bound (points: C) which is indicative of non-linearity in the data.

In conclusion we see that all tests proved positive for non-linearity, even though no visual evidence of non-linearity in the data plot was noticed.

(c) A plot of the overall residual sum of squares function, $S^2(\mu)$.

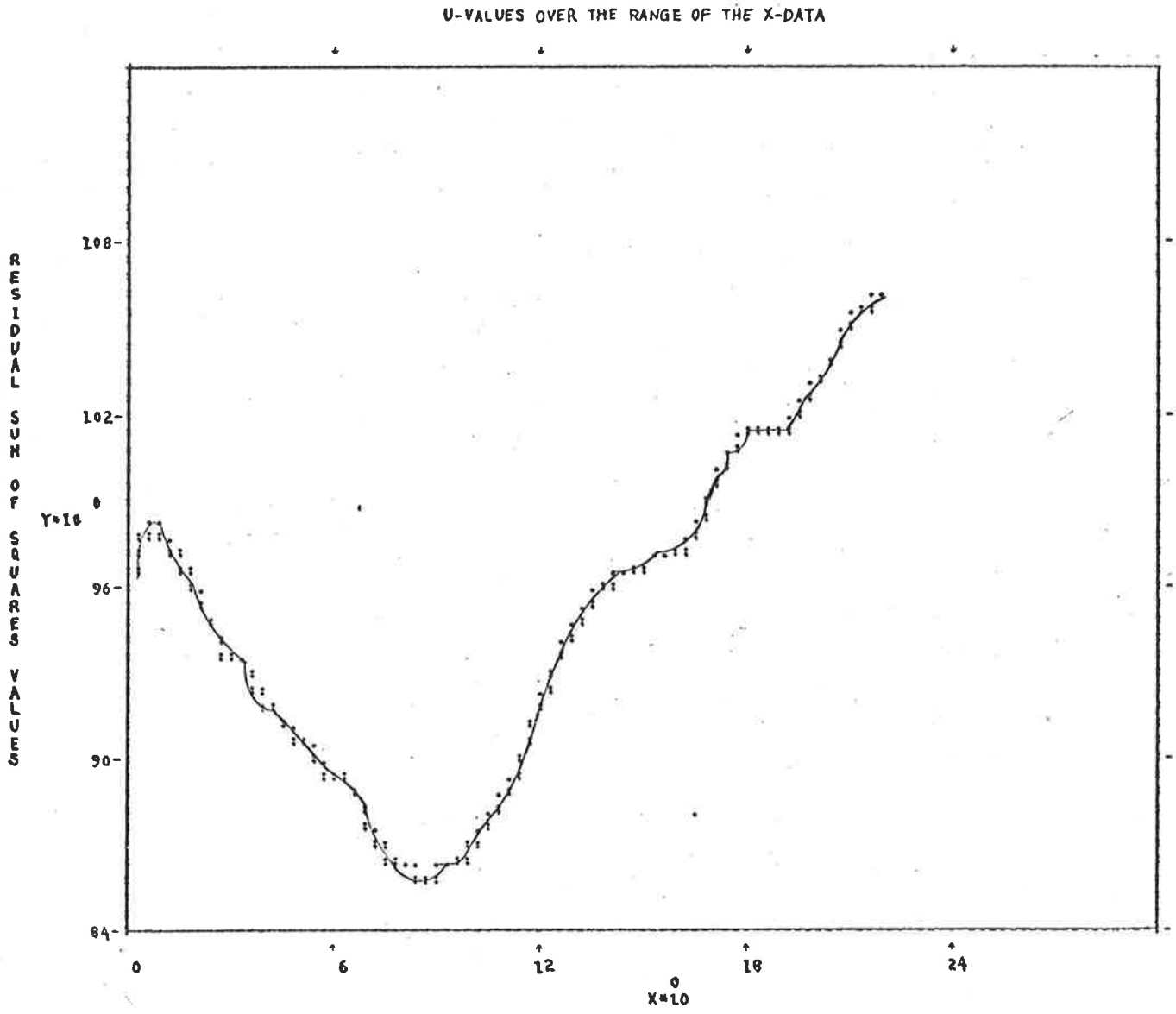


Fig. 5.5.(4)

(d)

Parameter	Estimate	Standard Error
α_1 (2.6)	2.02	.29
α_2 (5.0)	4.72	.51
β_1 (0.7)	0.81	.58
β_2 (0.5)	0.50	.32
γ (12)	8.89	1.29

and $\hat{\sigma}^2 = 0.899$ ($\sigma^2 = 1.00$).

In this case the least squares estimate of γ is a boundary point, $\hat{\gamma} = x_{41} = 8.99$.

Method		Time (sec $\times 10^{-2}$)
Hudson	§3.4	9.5
Hinkley	§3.5	7.6
Upper bound	§3.6	8.3
Interval	§3.7	6.5

An approximate 95 percent asymptotic confidence interval is given by (6.4, 11.4), which does not include the true value of γ(=12).

In conclusion, even though we can support the existence of a split-line model, the estimation procedure for γ becomes unreliable.

5.6 EXAMPLE SIX

(a) For equally spaced x values over the range from 0 to 22, 100 data points were generated for the split-line model.

$$E(y) = \begin{cases} 1 + x \\ 10.6 + 0.2x \end{cases} \quad \text{for } \gamma = 12,\ \sigma^2 = 1.$$

From the data we again obtain $\gamma = x_{55} = 12$ since the same x-values and intersection point (γ) are used as in the previous example.

In this particular case we have

1. $\eta_T|B| = 19.8$, which implies that the estimation procedure should be fairly reliable, according to Hinkley's criteria.
2. Non-linearity might be just perceivable from a plot, but it would be difficult to specify a fairly narrow sub-interval for $\hat{\gamma}$ with any confidence.

A Data Plot

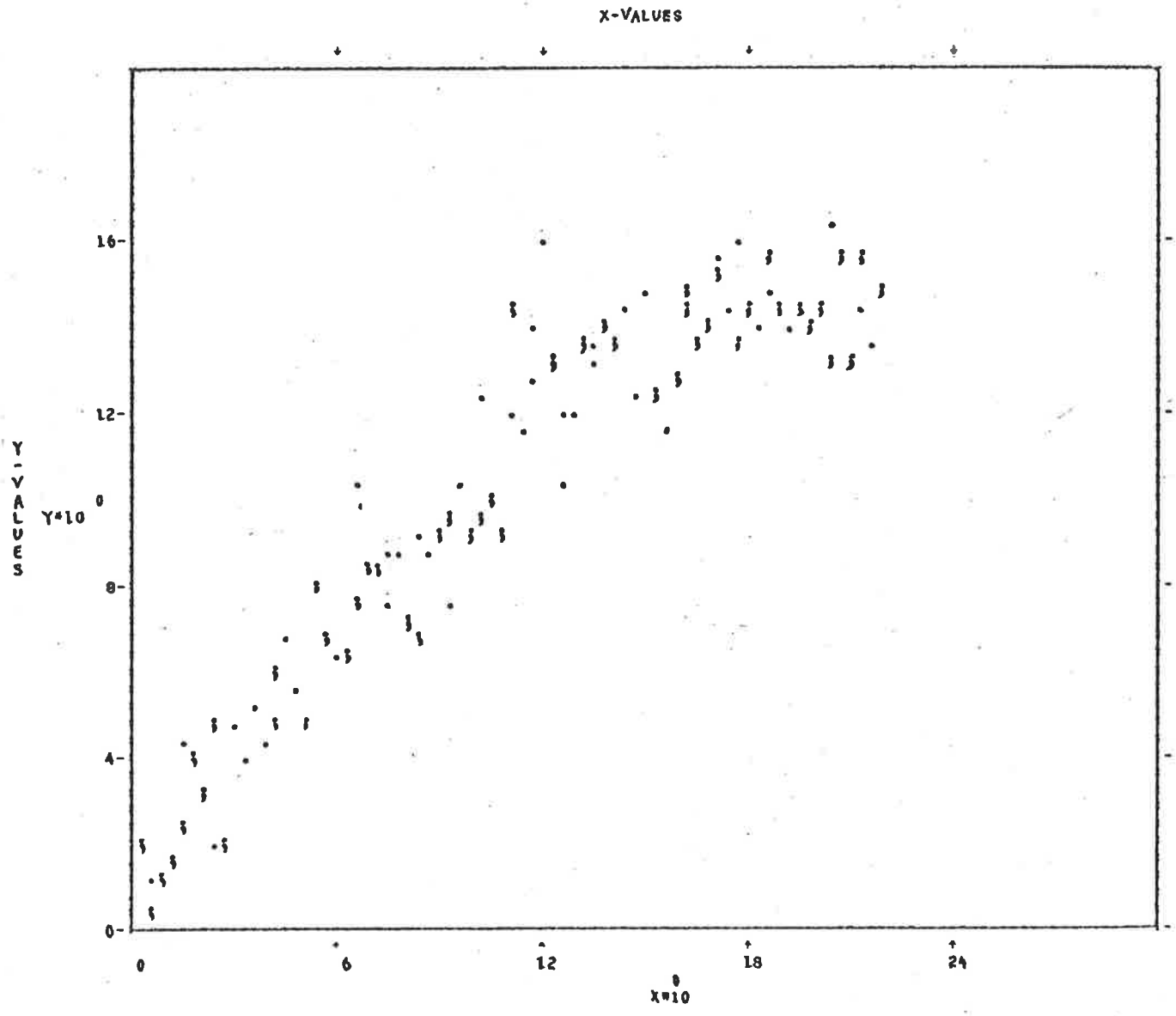


Fig. 5.6.(1)

(b) [1] Using differences between consecutive updated regression coefficients.

(i) Left to Right amalgamation.

A.O.V.

Comparisons	∂f	SS	MSS
Single line	1	1812.79	1812.79
Intersecting lines	1	51.31	51.31
Residual (y)	97	209.37	2.16
Total	99	2073.47	

The test statistic $t_1 = \frac{7.16}{\sqrt{2.16}} = 4.87$ gives a significant result for testing the non-linearity of the data.

(ii) Right to Left amalgamation.

A.O.V.

Comparisons	∂f	SS	MSS
Single line	1	1812.79	1812.79
Intersecting lines	1	85.14	85.14
Residual (y)	97	175.54	1.81
Total	99	2073.47	

Again the test statistic $t_1 = 6.86$ indicates a highly significant non-linear component in the data.

[2] The Quadratic Component technique.

A.O.V.

Comparisons	∂f	SS	MSS
Single line	1	1812.79	1812.79
Quadratic Component	1	122.12	122.12
Residual	97	138.56	1.43
Total	99	2073.47	

The test statistic $f = 122.12/1.43 = 85.49$ indicates the same result as above.

[3] The Graphic Approach: A plot of a 5 point moving Average for updated intercept values obtained by amalgamating from left to right.

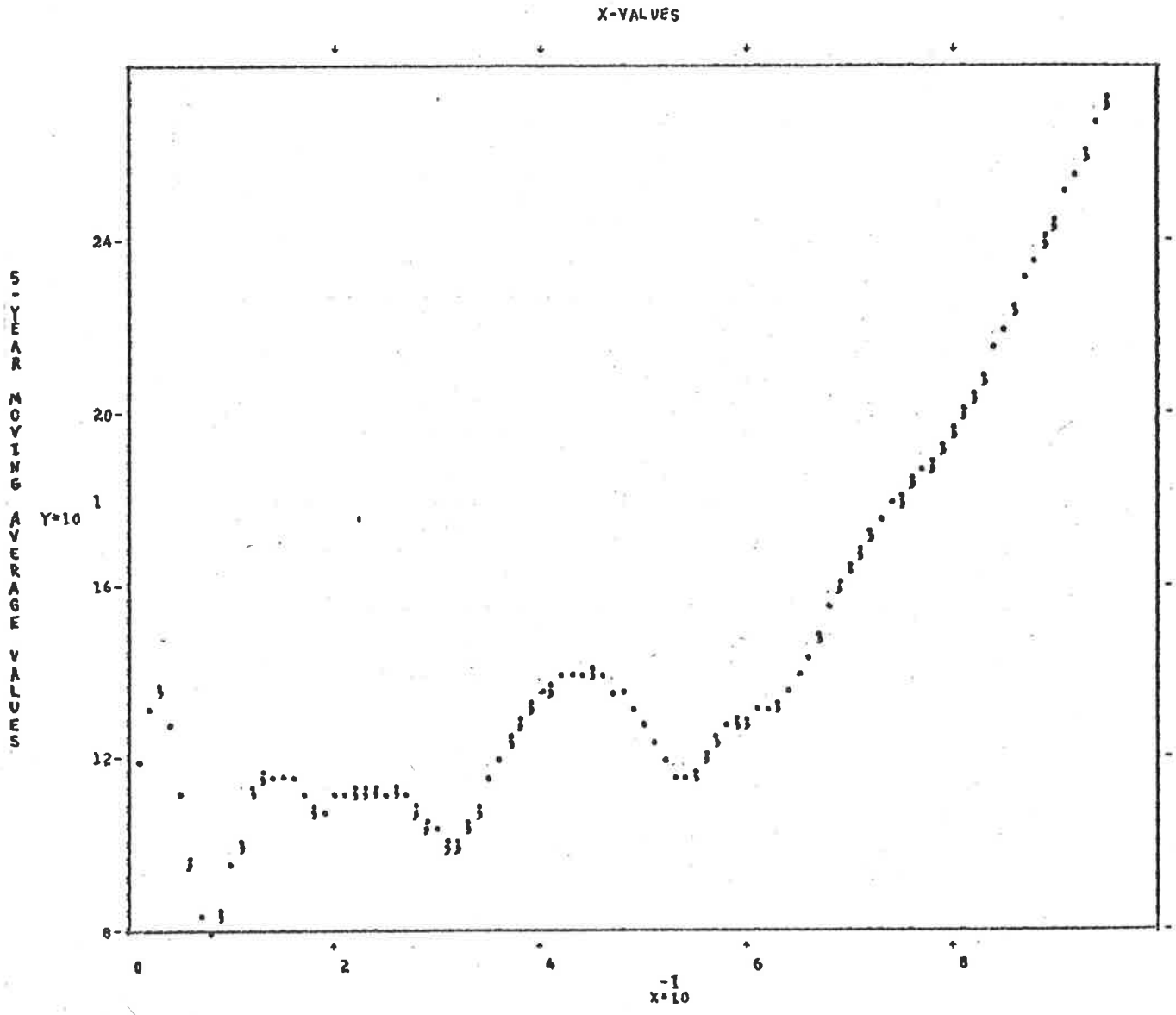


Fig. 5.6.(2)

The above indicates an even stronger non-linearity in the data than that of Fig. 5.5.(2).

[4] The Recursive Residual Approach, amalgamating from left to right.

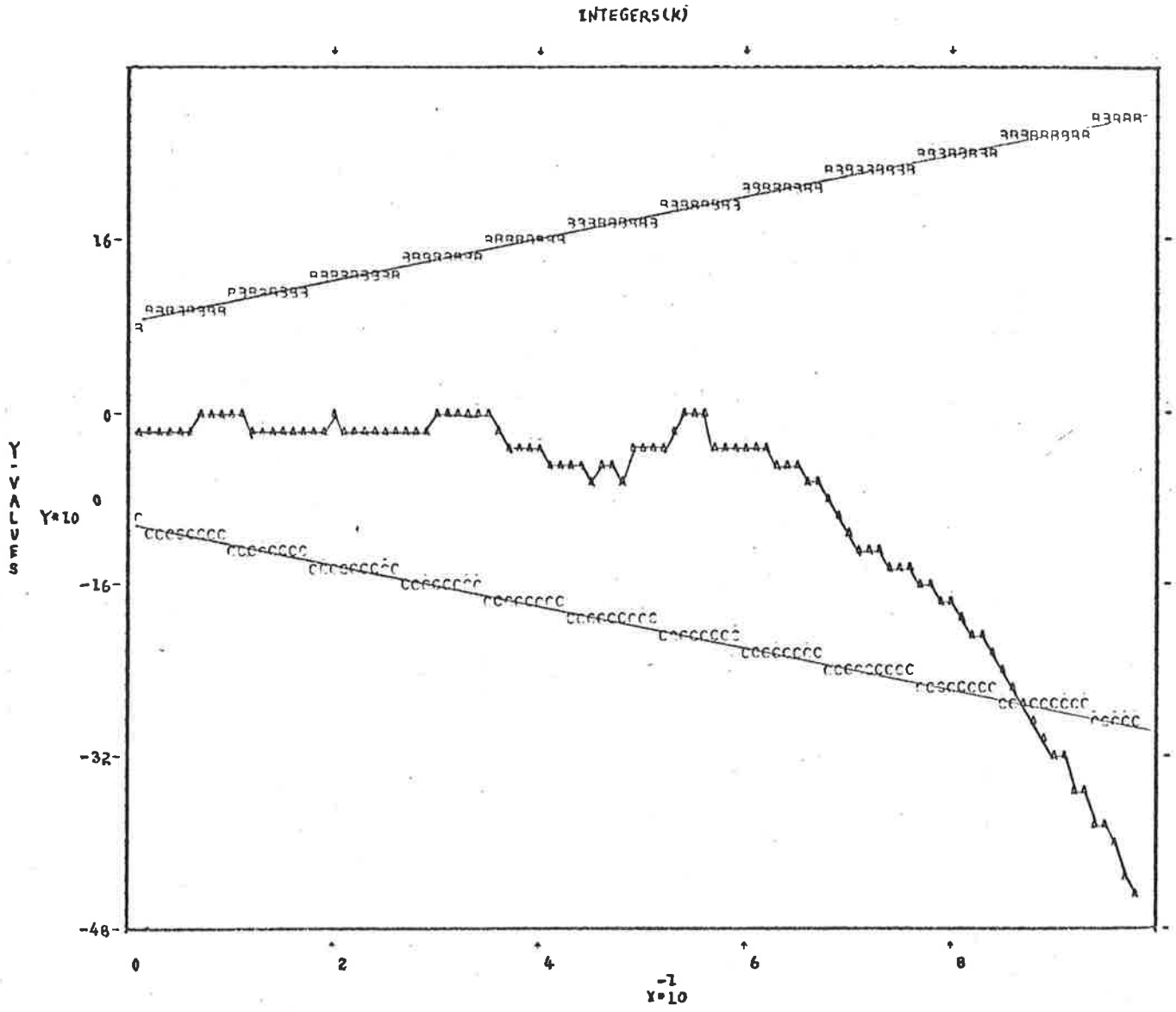


Fig. 5.6.(3)

Again, the above clearly demonstrates the significant non-linearity of the data.

(c) A plot of the overall Residual Sum of squares function, $S^2(\mu)$.

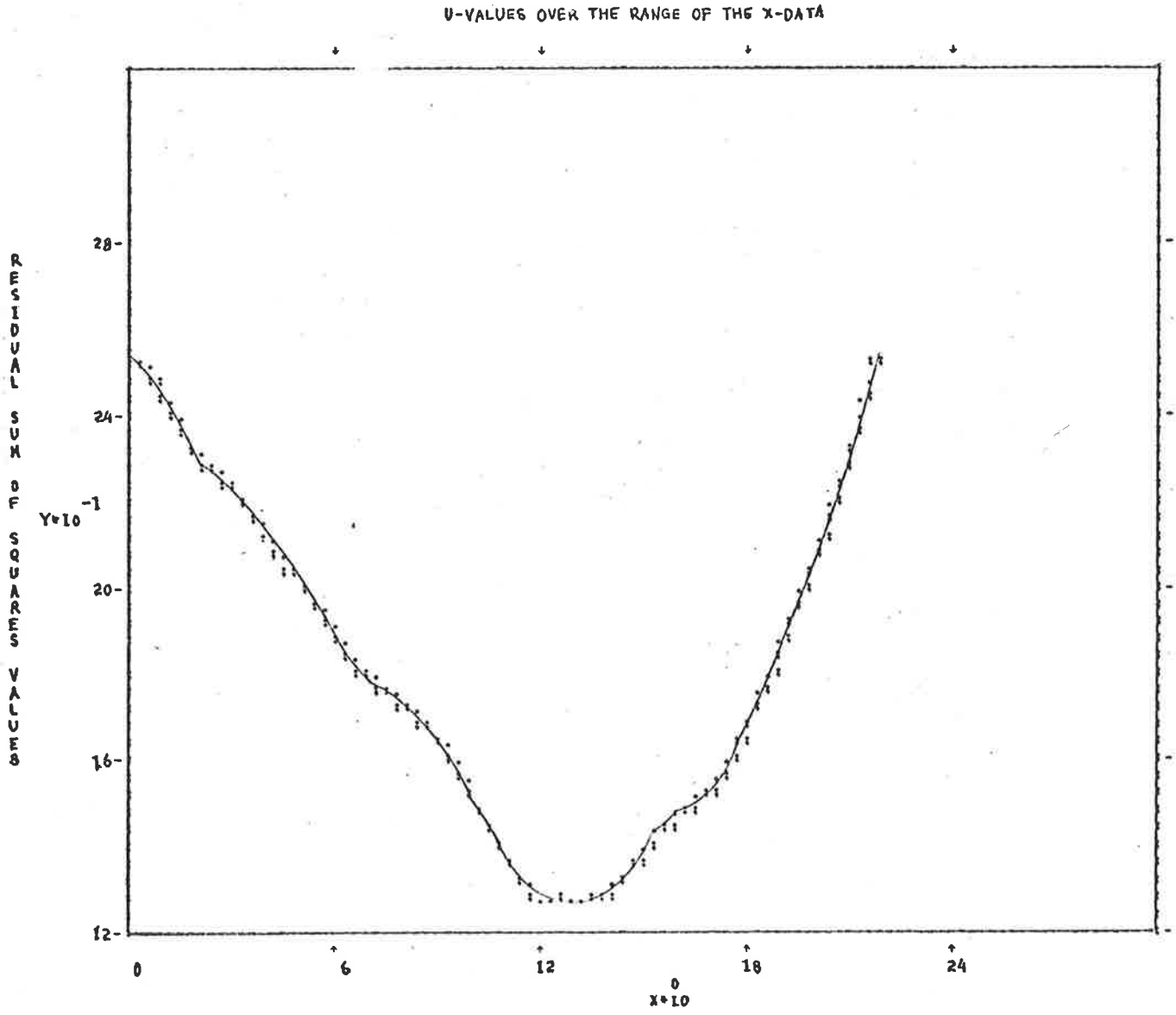


Fig. 5.6.(4)

(d)

Parameter	Estimate	Standard Error
α_1 (1.0)	1.13	.31
α_2 (10.6)	10.10	.11
β_1 (1.0)	9.66×10^{-1}	$.43 \times 10^{-1}$
β_2 (0.2)	2.36×10^{-1}	$.62 \times 10^{-1}$
γ (12)	12.28	.64

and $\hat{\sigma}^2 = 6.04 \times 10^{-1}$ $(\sigma^2 = 1)$.

In this case $\hat{\gamma}$ occurred at a smooth minimum and not at a boundary point as in example 5.5.

Method		Time (sec $\times 10^{-2}$)
Hudson	§3.4	10.6
Hinkley	§3.5	9.1
Upper bound	§3.6	8.3
Interval	§3.7	7.2

An approximate 95 percent asymptotic confidence interval is given by (11.02, 13.54).

5.7 EXAMPLE SEVEN

(a) On equally spaced x values over the range from 0 to 36 we generate 100 data points for the following model:

$$E(y) = \begin{cases} 17.8 - 1.1x: & \text{for } x \leqslant 12 \\ 5.8 - 0.1x: & \text{for } x \geqslant 12 \end{cases}$$

where $\gamma = 12$, $\sigma^2 = 1$ and $x_{34} < \gamma < x_{35}$. For this data we have

(i) $\eta_T|B| = 22.44$ thus the estimation procedure should be reliable.

(ii) the non-linearity in the data is visually apparent from a plot, and a sub-interval can be easily constructed in which $\hat{\gamma}$ is sure to be contained.

(iii) the intersection point (γ) is deliberately placed to one side of the range.

A Data Plot

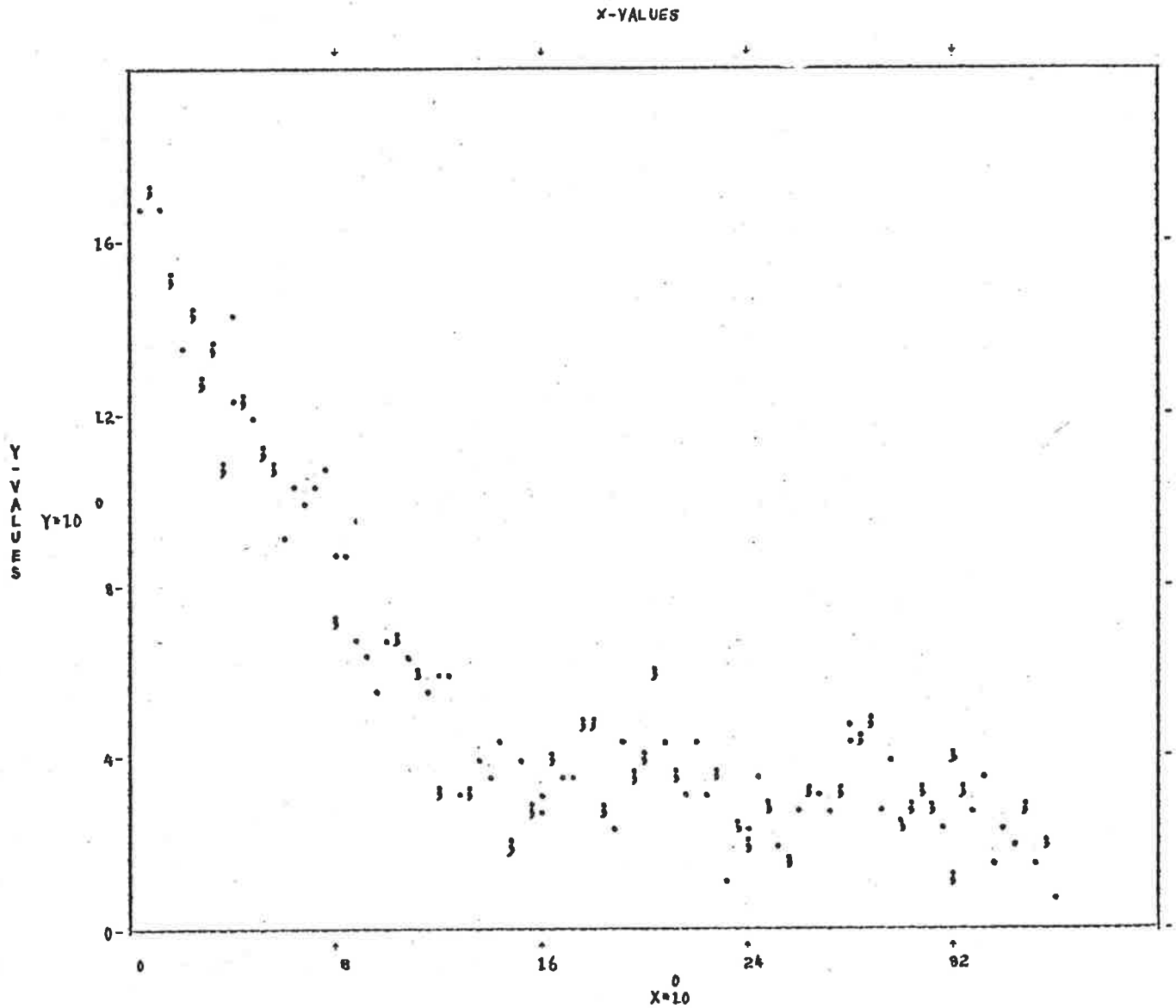


Fig. 5.7.(1)

(c) A plot of the overall residual sum of squares function, $S^2(\mu)$.

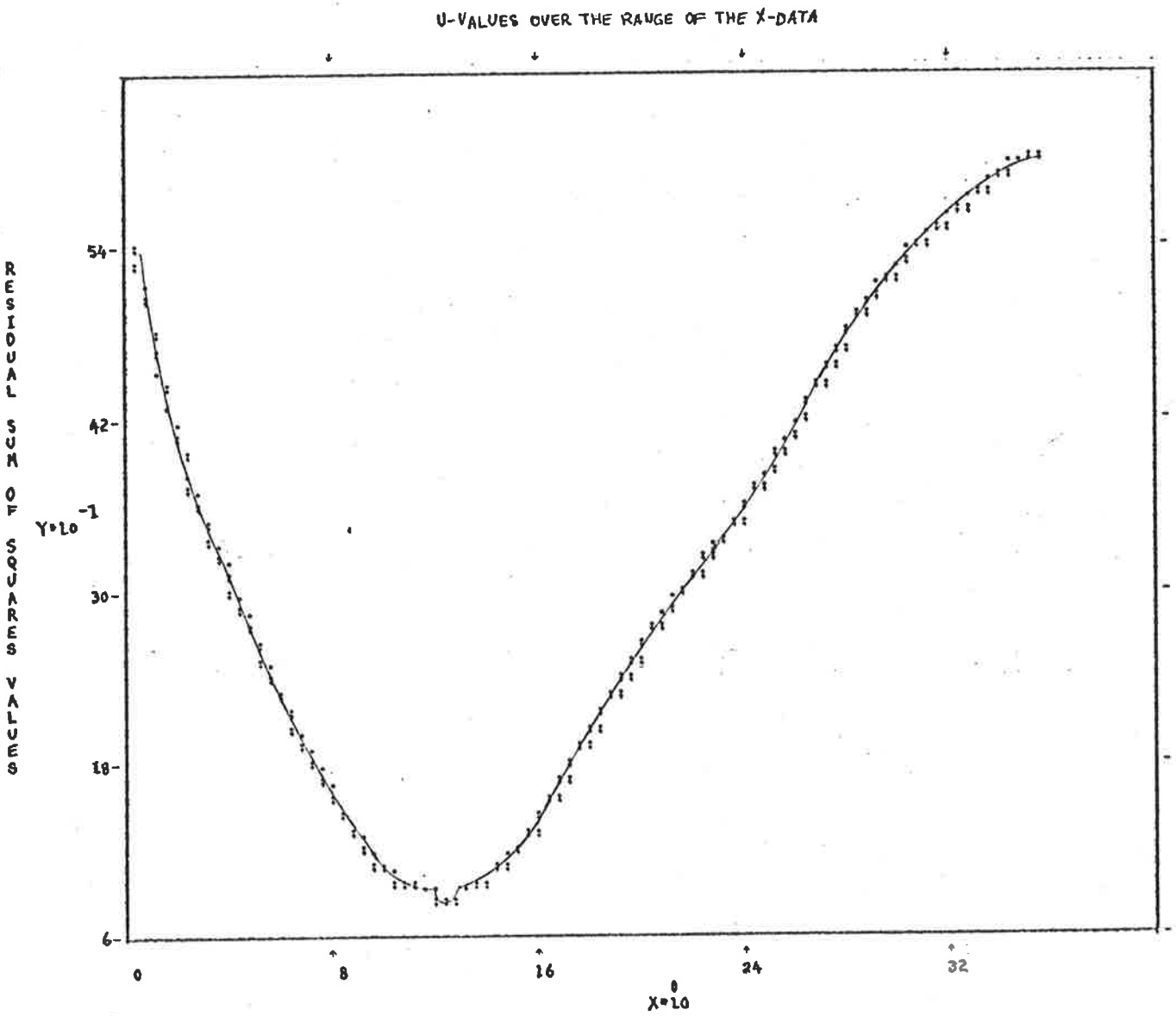


Fig. 5.7.(2)

(d)

Parameter	Estimate	Standard Error
α_1 (17.80)	17.52	.33
α_2 (5.80)	5.04	.46
β_1 (-1.10)	-1.06	$.46 \times 10^{-1}$
β_2 (-0.10)	-6.63×10^{-2}	$.18 \times 10^{-1}$
γ (12.00)	12.48	.42

and $\hat{\sigma}^2 = .989$ ($\sigma^2 = 1.00$) with $\hat{\gamma}$ lying between the x values (12.36, 12.73); e.g. $x_{35} < \hat{\gamma} < x_{36}$.

Method		Time (sec $\times 10^{-2}$)
Hudson	§3.4	9.3
Hinkley	§3.5	8.3
Upper bound	§3.6	7.8
Interval	§3.7	7.0

Since non-linearity is visually apparent from the plot (Fig. 5.7.(1)) we choose $\ell_1 = 23$, $\ell_2 = 45$ such that we obtain the initial interval (8.00, 16.00), which gives the following times

Method		Time (sec $\times 10^{-2}$)
Upper bound	§3.6	7.8
Interval	§3.7	6.6

An approximate 95 percent asymptotic confidence interval is given by (11.67, 13.30).

5.8 EXAMPLE EIGHT

(a) For equally spaced x values over the range 0 to 16 we generate 100 data points for the following split-line model:

$$E(y) = \begin{cases} 5 + 2x: & x \leqslant 12 \\ 41 - x : & x \geqslant 12 \end{cases}$$

where $\gamma = 12$, $\sigma^2 = 1$ and $x_{75} < \gamma < x_{76}$.

Under Hinkley's criteria, the estimation procedure should be reliable. The interest in this data is that the intersection point is clearly indicated on the data plot and the estimate of γ can be narrowed down to a small subinterval.

A Data Plot

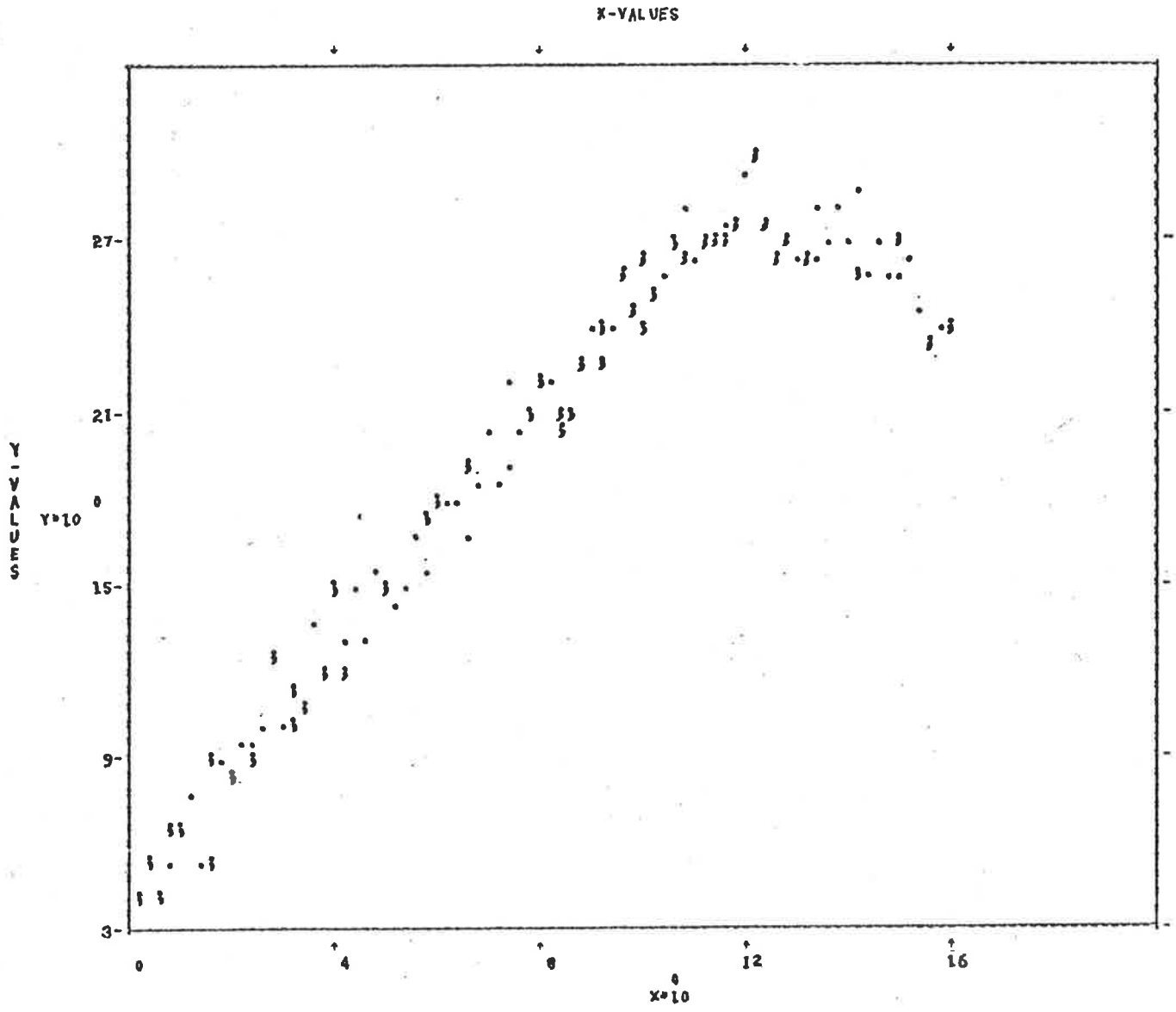


Fig. 5.8.(1)

(c) A Plot of the overall residual sum of squares function, $S^2(\mu)$.

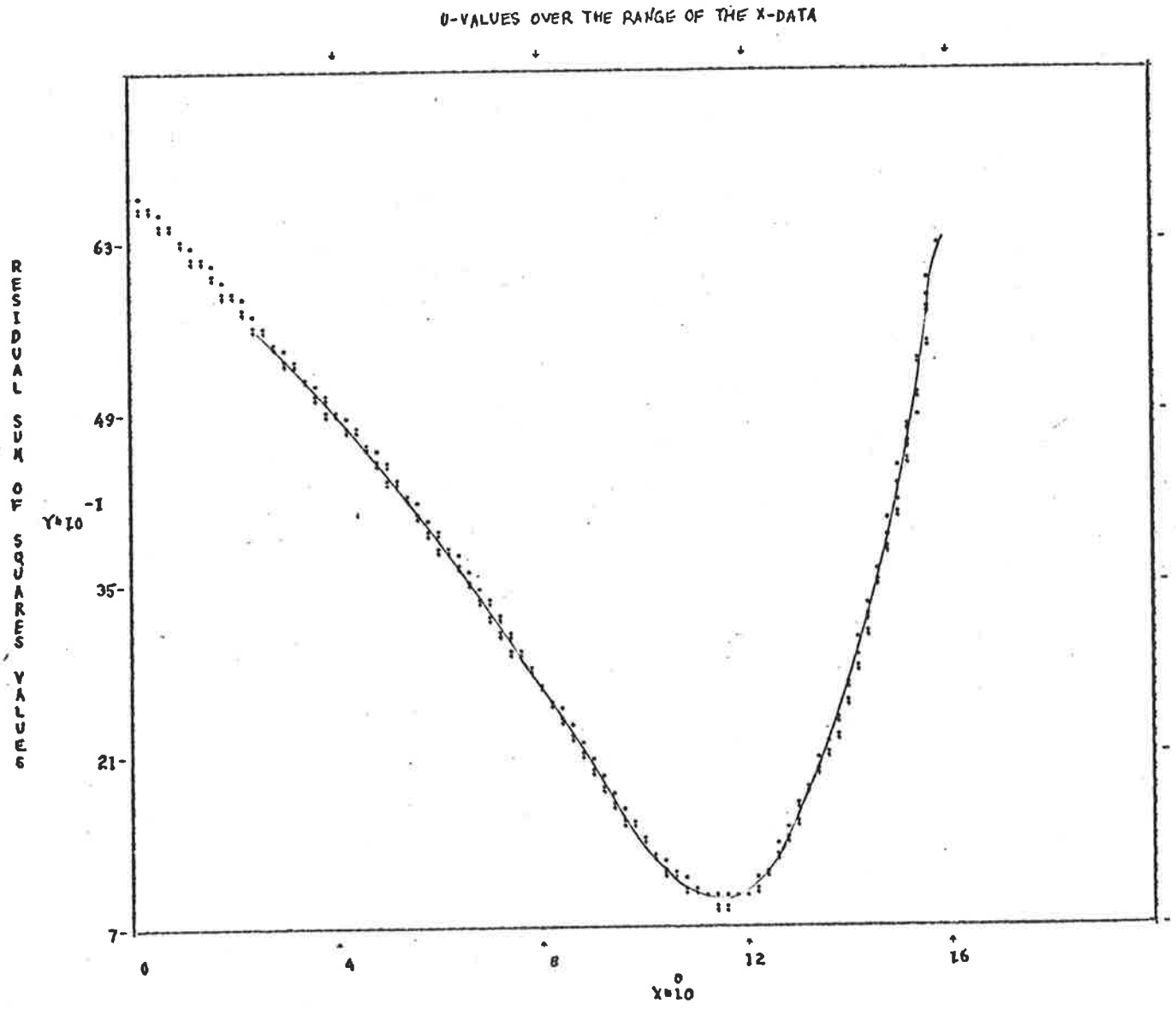


Fig. 5.8.(2)

(d)

Parameter	Estimate	Standard Error
α_1 (5.00)	4.65	.23
α_2 (41.00)	39.21	2.15
β_1 (2.00)	2.07	.35 $\times 10^{-1}$
β_2 (-1.00)	-0.89	.15
γ (12.00)	11.69	.15

and $\hat{\sigma}^2 = 1.02$ ($\sigma^2 = 1$), with $\hat{\gamma}$ lying between the consecutive x-values (11.64, 11.8); e.g. $x_{74} < \hat{\gamma} < x_{75}$.

Method		Time (sec $\times 10^{-2}$)
Hudson	§3.4	11.10
Hinkley	§3.5	9.4
Upper bound	§3.6	8.5
Interval	§3.7	7.4

Choosing $\ell_1 = 63$, $\ell_2 = 85$ we obtain an interval (10.02, 13.58), which gives the following computational times

Method		Time (sec $\times 10^{-2}$)
Upper bound	§3.6	8.0
Interval	§3.7	6.3

An approximate 95 percent asymptotic confidence interval is given by (11.39, 12.00).

5.9 EXAMPLE NINE

(a) To demonstrate the efficiencies of each estimation method for large samples we generated 360 values within the range (0, 30). The intersection point was chosen to lie to the extreme right of the range; the split-line model being,

$$E(y) = \begin{cases} x & : \quad x \leqslant 25 \\ 50 - x & : \quad x \geqslant 25 \end{cases}$$

where $\gamma = 25$, $\sigma^2 = 4$ and $x_{300} < \gamma < x_{301}$.

This data should ensure reliable estimation procedures, and from a data plot we see that non-linearity is barely discernible on the extreme right.

A Data Plot

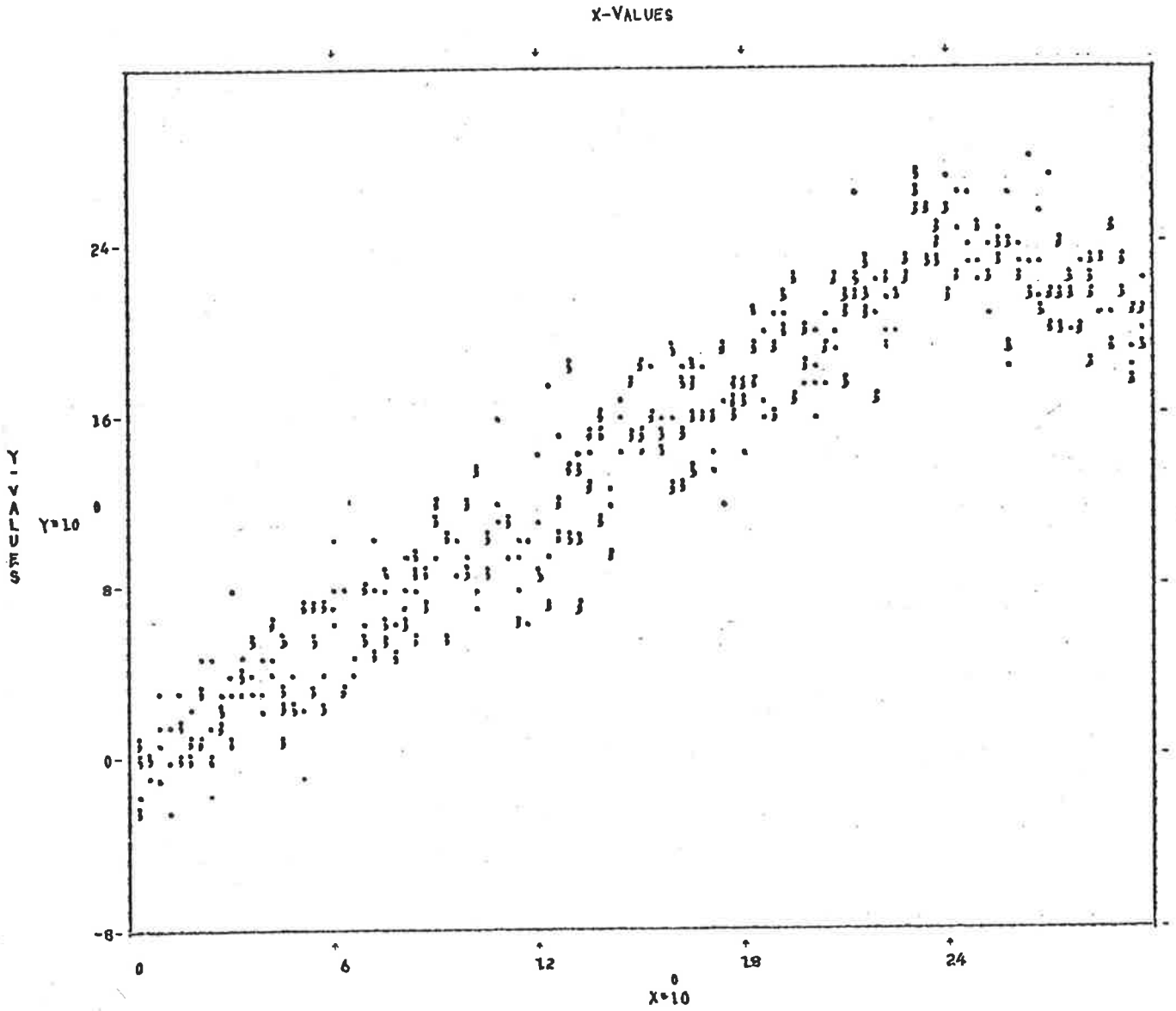


Fig. 5.9.(1)

(b) [3] The Graphic Approach demonstrates how the moving average plot tends to level out as we proceed through enough values of one regression line and begins, in this case, an upward swing as the data from the second regression line begins to influence the updated estimates for the intercept.

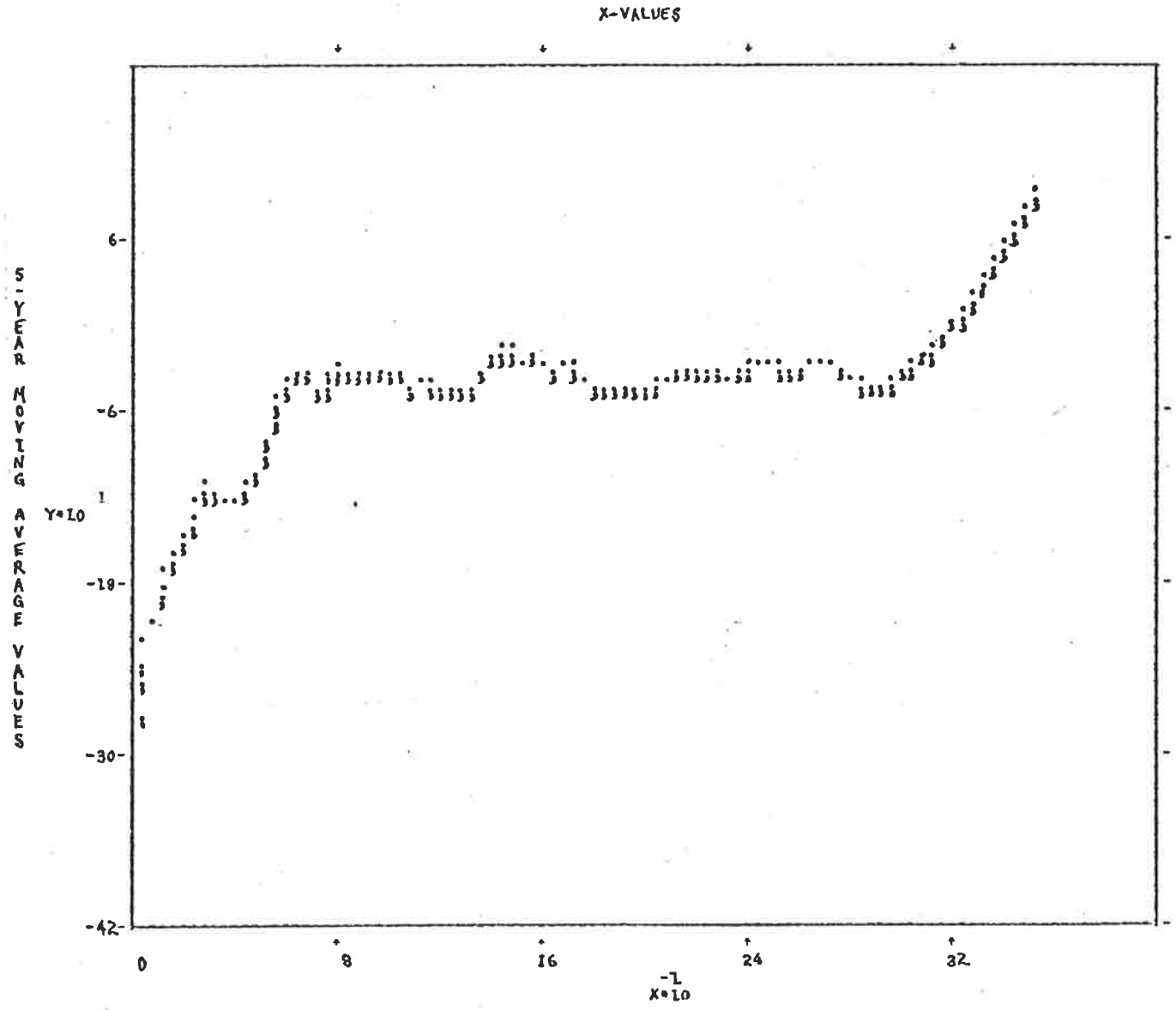


Fig. 5.9.(2)

(c) A plot of the overall residual sum of squares function, $S^2(\mu)$.

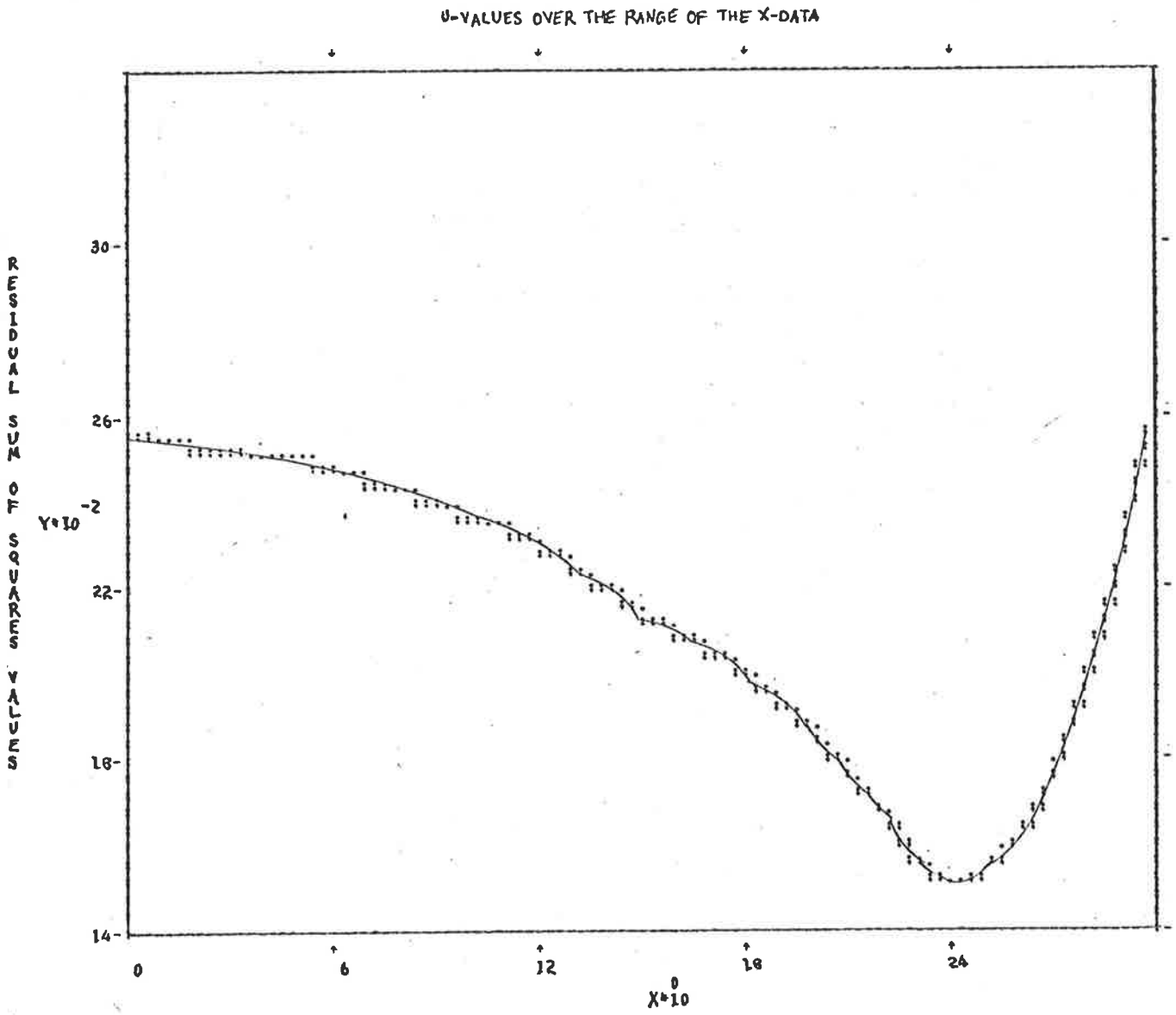


Fig. 5.9.(3)

(d)

Parameter	Estimate	Standard Error
α_1 (0.00)	-3.69×10^{-1}	.24
α_2 (50.00)	40.93	4.18
β_1 (1.00)	1.02	$.17 \times 10^{-1}$
β_2 (-1.00)	-0.67	.15
γ (25.00)	24.39	.33

and $\hat{\sigma}^2 = 4.29$, $(\sigma^2=4)$ with $\hat{\gamma}$ lying between the consecutive x-values (24.32, 24.40); e.g. $x_{292} < \hat{\gamma} < x_{293}$.

Method		Time (sec $\times 10^{-2}$)
Hudson	§3.4	35.4
Hinkley	§3.5	27.8
Upper bound	§3.6	24.6
Interval	§3.7	17.9

Using $\ell_1 = 212$, $\ell_2 = 313$: which gives an initial interval of (19.97, 25.99), we obtain the following computational times,

Method		Time (sec $\times 10^{-2}$)
Upper bound	§3.6	22.9
Interval	§3.7	13.9

An approximate 95 percent asymptotic confidence interval is given by (23.74, 25.02).

5.10 Conclusions

In the preceding and other examples analysed, several patterns were observed when comparing the time taken between methods in finding the estimate of the intersection point, γ. Firstly in all cases analysed, Hinkley's approach (§3.5) was computational faster than Hudson's method (§3.4). This was predictable, due to the reasons given in §3.5.

The Interval Estimation method of §3.7, although not guaranteeing the least squares estimate of γ, did in fact produce the least squares estimate in all cases. The generated data sets given here were selected to demonstrate the versatility of this method, under the greatly varying conditions demonstrated by the data plots. The effectiveness of §3.7 is shown by these examples, which range from data plots with no visual evidence of non-linearity to data plots with unmistakable non-linearity; with γ being positioned throughout the range of the data. In all the cases considered, for $M \geq 30$, method §3.7 was seen to be the quickest; with the relative efficiency of this method increasing as M increases (re 5.9). Further saving in computational time was observed (for data with $M \geq 30$) if, from a data plot, we can obtain an initial interval (see §3.7) with which to begin this method.

In the cases where $M < 30$, for method §3.7, it was found that the process of finding the interval (in which to search for an estimate of γ - see §3.7) consumed too much time and consequently this method took longer to find an estimate than the other methods. Yet for $20 \leq M < 30$ the method of §3.7, although slower than §3.5 and §3.6, was in fact faster than §3.4.

A summary as to which method to apply can be given as follows:-

Size (M)	Method
$M < 20$	Hinkley's (§3.5) method - produces the least squares estimate of γ.
$20 \leqslant M < 30$	§3.5 or §3.6 - produces the least squares estimate of γ.
$M \geqslant 30$	§3.6 - the quickest method which guarantees the least squares estimate of γ. If we are willing to forego this guarantee the Interval Approach of §3.7 is significantly faster, with the relative saving in computational time increasing as M increases. It seems reasonable to use §3.7 because 1. of the real saving in time. 2. in all cases analysed, the least squares estimate of γ was found.

APPENDIX

The following is a listing of the subroutines and functions for finding an estimate of the intersection point, γ, in situations where the split-line regression model is appropriate.

This listing, in FORTRAN IV, implements the procedures described in sections 3.4, 3.5, 3.6 and 3.7.

These subroutines and functions have been tested on the University of Adelaide CDC 6400, under the operating system SCOPE 4.6, using the FTN compiler.

A.

```
      SUBROUTINE HUDSON (X,Y,NI,N,M,CSY)

      COMMON/A/XSUM(360),YSUM(360),XYSUM(360),X2SUM(360)
      COMMON/B/X1,X2,Y1,Y2,C1,C2
      COMMON/F/J1,J2,I2
      COMMON/D/B1,B2,G
      DIMENSION X(1) , Y(1) , NI(1)
      INTEGER T


C     THIS IS METHOD 3.2.....................................................
C     HUDSONS EXHAUSTIVE SEARCHING PROCEDURE TO FIND THE LEAST SQUARES
C     ESTIMATE FOR THE INTERSECTION POINT , GAMMA , BY MINIMIZING THE
C     RESIDUAL SUM OF SQUARES FUNCTION AS GIVEN BY EQUATION 3.2.(E)

C     ROUTINES CALLED BEFORE ENTRY INTO HUDSON- SORT , ACUM
C     ROUTINES CALLED WITHIN HUDSON - WORK , FUNC

C     INPUT VALUES -
C        X : ARRAY OF INCREASING X VALUES
C        Y : ARRAY OF CORRESPONDING Y VALUES
C        NI,N,M,-AS DEFINED IN ROUTINE ACUM

C     IMPORTANT COMPUTED VALUES -
C        GHAT :THE FINAL ESTIMATE FOR THE INTERSECTION POINT
C           T :THE PARTITIONING FROM WHICH THE ABOVE ESTIMATE IS FOUND
C           S:ESTIMATE FOR THE VARIANCE OF THE SPLIT-LINE MODEL
C          NN :THE NUMBER OF PARTITIONS BEING SEARCHED

      NN = M-2

C     SET THE INITIAL VALUES FOR THE SEARCHING PROCEDURE
C     STEP 1.................................................................
      T = 1
      GHAT = X(1)
      RESS = CSY-(XYSUM(M)-XSUM(M)*YSUM(M)/N)**2/(X2SUM(M)-XSUM(M)**2/N)
      DO 40 I = 2,NN
C     STEP TWO...............................................................
      CALL WORK(I,1,N,M,NI)
      CALL FUNC(G,N,M,J1,I,S1,AA1,BB1,BB2,CSY,0)
      IF(S1.GE.RESS) GO TO 40
      IF(G.LT.X(J1).OR.G.GT.X(I2)) GO TO 10
      RESS = S1
      T = I
      GHAT = G
      GO TO 40
C     STEP THREE.............................................................
   10 IF(G.LT.X1.OR.G.GT.X2) GO TO 20
      GX = X(J1)
      IF(G.GT.X(I2)) GX = X(I2)
```

```
      CALL FUNC(GX,N,M,J1,I,S1,AA1,BB1,BB2,CSY,0)
      IF(S1.GE.RESS) GO TO 40
      RESS = S1
      T = I
      GHAT = GX
      GO TO 40
C     STEP FOUR.....................................................
   20 CALL FUNC(X(J1),N,M,J1,I,SX1,AA1,BB1,BB2,CSY,0)
      CALL FUNC(X(I2),N,M,J1,I,SX2,AA1,BB1,BB2,CSY,0)
      GX = X(J1)
      SX = SX1
      IF(SX1-SX2.LE.1.E-20) GO TO 30
      GX = X(I2)
      SX = SX2
   30 IF(SX.GE.RESS) GO TO 40
      RESS = SX
      T = I
      GHAT = GX
   40 CONTINUE
C     AT THIS STAGE , THE SEARCHING PROCEDURE HAS ENDED.............
      CALL WORK(T,0,N,M,NI)
      XL = X(J1)
      XU = X(I2)
C     XL = X(J1) ≤ GHAT ≤ X(I2) = XU
      CALL FUNC(GHAT,N,M,J1,T,S,AA1,BB1,BB2,CSY,1)
      AA2 = AA1 + GHAT*(BB1-BB2)
      S = RESS / (N - 4)
      CNAME = 10H    HUDSON
      PRINT 50,CNAME,J1,XL,XU,GHAT,AA1,AA2,BB1,BB2,S
   50 FORMAT(1H1/10X*THE PROCEDURE USED IS *A10/10X*THE NUMBER OF Y-VA_U
     ZES FOR WHICH X≤GHAT =*I3/10X*THE IMMEDIATE X-VALUES BOUNDING GHAT
     ZARE*E10.4*,*E10.4/10X*THE ESTIMATE FOR THE INTERSECTION POINT,GHAT
     Z =*E10.4/10X*THE ESTIMATES OF ALPHA1,ALPHA2,BETA1,BETA2 ARE *3(E10
     Z.4*,*),E10.4/10X*ESTIMATE OF THE VARIANCE FOR THE SPLIT-LINE MODEL
     Z = *E10.4)
      RETURN
      END
```

B.

```
      SUBROUTINE HINKLEY (X,Y,NI,N,M,CSY)
      COMMON/A/XSUM(360),YSUM(360),XYSUM(360),X2SUM(360)
      COMMON/B/X1,X2,Y1,Y2,C1,C2
      COMMON/D/B1,B2,G
      COMMON/E/CT,DT,ET
      COMMON/F/J1,J2,I2
      DIMENSION X(1) , Y(1) , NI(1)
      INTEGER T

C     THIS IS METHOD 3.5 .
C     HINKLEYS EXHAUSTIVE SEARCHING PROCEDURE TO FIND THE LEAST SQUARES
C     ESTIMATE FOR THE INTERSECTION POINT , GAMMA, BY MAXIMIZATION OF A
C     SUBFUNCTION OF THE RESIDUAL SUM  OF  SQUARES FUNCTION AS GIVEN BY
C     EQUATION 3.2.(F)

C     INPUT VALUES :
C      X :ARRAY OF INCREASING X-VALUES
C      Y : CORRESPONDING ARRAY OF Y-VALUES
C N,M,NI :AS DEFINED IN SUBROUTINE ACUM

C     SUBROUTINES TO BE CALLED BEFORE ENTRY INTO THIS ROUTINE -
C          ACUM, AND POSSIBLY SORT.

C     NN:THE NUMBER OF PARTITIONS TO SEARCH .
      NN = M-2

      CSX = X2SUM(M) - XSUM(M)**2 / N
C     SET THE INITIAL VALUE OF Z TO ZERO .
C     STEP 1 ..........................................................
      T = 1
      GHAT = X(1)
      Z = 0
      DO 40 I = 2,NN
C     STEP TWO.........................................................
      CALL WORK(I,3,N,M,NI)
      Z1 = ZT (G,G,CSX,B1,B2,CT,DT,ET)
      IF (Z1.LE.Z) GO TO 40
      IF(X(J1).GT.G.OR.G.GT.X(I2)) GO TO 10
      Z = Z1
      T = I
      GHAT = G
      GO TO 40
C     STEP THREE.......................................................
   10 DF = DT - G*ET
      DG = (CT - G*DT) / DF
      IF(DF.GT.0.AND.DG.LT.X(I2).OR.DF.LT.0.AND.X(J1).LT.DG) GO TO 20
      GX = X(J1)
      IF(G.GT.X(I2)) GX = X(I2)
      ZX = ZT(GX,G,CSX,B1,B2,CT,DT,ET)
      IF(ZX.LE.Z) GO TO 40
```

```
      Z = ZX
      T = I
      GHAT = GX
      GO TO 40
C     STEP FOUR.....................................................
   20 ZX1 = ZT(X(J1),G,CSX,B1,B2,CT,DT,ET)
      ZX2 = ZT(X(I2),G,CSX,B1,B2,CT,DT,ET)
      GX = X(J1)
      ZX = ZX1
      IF(ZX1-ZX2.GT.1.E-20) GO TO 30
      GX = X(I2)
      ZX = ZX2
   30 IF(ZX.LE.Z) GO TO 40
      Z = ZX
      T = I
      GHAT = GX
   40 CONTINUE
C     THE SEARCHING PROCEDURE ENDS AT THIS STAGE .
      CNAME = 10H   HINKLEY
      CALL WORK(T,0,N,M,N1)
      CALL FUNC(GHAT,N,M,J1,T,RESS,AA1,BB1,BB2,CSY,1)
      XL = X(J1)
      XU = X(I2)
      AA2 = AA1 + GHAT*(BB1-BB2)
      S = (CSY- (XYSUM(M)-XSUM(M)*YSUM(M) / N )**2 / CSX - Z) / (N-4)
      PRINT 50,CNAME,J1,XL,XU,GHAT,AA1,AA2,BB1,BB2,S
   50 FORMAT(1H1/10X*THE PROCEDURE USED IS *A10/10X*THE NUMBER OF Y-VA_U
     ZES FOR WHICH X≤GHAT =*I3/10X*THE IMMEDIATE X-VALUES BOUNDING GHAT
     ZARE*E10.4*,*E10.4/10X*THE ESTIMATE FOR THE INTERSECTION POINT,GHAT
     Z =*E10.4/10X*THE ESTIMATES OF ALPHA1,ALPHA2,BETA1,BETA2 ARE *3(E10
     Z.4*,*),E10.4/10X*ESTIMATE OF THE VARIANCE FOR THE SPLIT-LINE MODEL
     Z = *E10.4)
      RETURN
      END
```

C.

```
      SUBROUTINE UBOUND(X,Y,NI,N,M,CSY)
      COMMON/A/XSUM(360),YSUM(360),XYSUM(360),X2SUM(360)
      COMMON/B/X1,X2,Y1,Y2,C1,C2
      COMMON/C/INTER,IA,IB
      COMMON/D/B1,B2,G
      COMMON/E/CT,DT,ET
      COMMON/F/J1,J2,I2
      DIMENSION X(1) , Y(1) , NI(1)
      INTEGER T
      CSX = X2SUM(M) - XSUM(M)**2 / N

C     THIS IS METHOD 3.6 WHERE THE LEAST SQUARES ESTIMATE OF GAMMA IS
C     FOUND THE USE OF AN UPPER BOUND FOR THE RESIDUAL SUM OF SQUARES .
C     THIS IS ACHIEVED BY WORKING WITH THE LOWER BOUND FOR THE SUB-
C     FUNCTION ( Z ) AS GIVEN IN EQUATION 3.2.(F)

C     VALUES OF IMPORTANCE-
C     INTER: (1)-FOR INTER = 0 IMPLIES NO PLOT OF THE DATA IS AVAILABLE
C                 TO OBTAIN A REGION IN WHICH IT IS SUSPECTED THAT GAMMA
C                 MAY LIE .
C            (2)-IF INTER = 1 THEN FROM A PLOT STUDY WE OBTAIN L1 , L2
C                 WHERE L1:IS THE POSITION OF THAT VALUE OF X JUST LESS
C                          THAN THE INITIAL LOWER LIMIT OF THE INTERVAL .
C                       L2: WHERE X(L2-1)≤ B ≤ X(L2),B BEING THE UPPER
C                           BOUNDARY POINT,WITH 1≤ L2 ≤ L2 ≤ M .

C     ROUTINES CALLED BEFORE ENTRY INTO UBOUND - SORT , ACUM
C     ROUTINES CALLED WITHIN UBOUND - WORK , FUNC
C     FUNCTIONS CALLED WITHIN UBOUND- ZT

C     THE INPUT VALUES X,Y,NI,N,M,CSY ARE DEFINED AS IN SUBROUTINE
C     HINKLEY

C     STEP 1: CALCULATING THE INITIAL UPPER BOUND .
      L1 = M / 3 + 1
      L2 = M - L1
      IF(INTER.EQ.0) GO TO 10
      L1 = IA
      L2 = IB
   10 T = (L1 + L2) / 2
      CALL WORK(T,3,N,M,NI)
      IF(X(J1).GT.G.OR.G.GT.X(I2)) GO TO 30
      ZBOUND = ZT(G,G,CSX,B1,B2,CT,DT,ET)
      GHAT = G
      GO TO 50
   30 ZX1 = ZT(X(J1),G,CSX,B1,B2,CT,DT,ET)
      ZX2 = ZT(X(I2),G,CSX,B1,B2,CT,DT,ET)
      ZX = ZX1
      GX = X(J1)
      IF(ZX1-ZX2.GT.1.E-20) GO TO 40
      ZX = ZX2
      GX = X(I2)
   40 ZBOUND = ZX
      GHAT = GX
   50 CONTINUE
```

```
C       THE SEARCHING PROCEDURE FOR FINDING THE LEAST SQUARES ESTIMATE
C       OF THE INTERSECTION POINT GAMMA , USING METHOD 3.5  .
C       STEP 2: THE SEARCHING BEGINS .

        NN = M-2
        DO 90 I = 2,NN
        CALL WORK(I,3,N,M,NI)
        Z1 = ZT(G,G,CSX,B1,B2,CT,DT,ET)
        IF(ZBOUND.GE.Z1) GO TO 90
        IF(X(J1).GT.G.OR.X(I2).LT.G) GO TO 60
        ZBOUND = Z1
        T = I
        GHAT = G
        GO TO 90
   60   DF = DT - G*ET
        DG = (CT - G*DT) / DF
        IF(DF.GT.0.AND.DG.LT.X(I2).OR.DF.LT.0.AND.X(J1).LT.DG) GO TO 70
        GX = X(J1)
        IF(G.GT.X(I2))GX = X(I2)
        ZX = ZT(GX,G,CSX,B1,B2,CT,DT,ET)
        IF(ZX.LE.ZBOUND) GO TO 90
        ZBOUND = ZX
        T = I
        GHAT = GX
        GO TO 90
   70   ZX1 = ZT(X(J1),G,CSX,B1,B2,CT,DT,ET)
        ZX2 = ZT(X(I2),G,CSX,B1,B2,CT,DT,ET)
        GX = X(J1)
        ZX = ZX1
        IF(ZX1-ZX2.GT.1.E-20) GO TO 80
        GX = X(I2)
        ZX = ZX2
   80   IF(ZBOUND.GT.ZX) GO TO 90
        ZBOUND = ZX
        T = I
        GHAT = GX
   90   CONTINUE
C       THE SEARCHING PROCEDURE ENDS .
        CALL WORK(T,0,N,M,NI)
        XL = X(J1)
        XU = X(I2)
        CALL FUNC(GHAT,N,M,J1,T,RSS,AA1,BB1,BB2,CSY,1)
        AA2 = AA1 + GHAT*(BB1-BB2)
        S= (CSY- (XYSUM(M)-XSUM(M)*YSUM(M)/N)**2 / CSX - ZBOUND) / (N-4)
        CNAME = 10H    SBOUND
        PRINT 100,CNAME,J1,XL,XU,GHAT,AA1,AA2,BB1,BB2,S
  100   FORMAT(1H1/10X*THE PROCEDURE USED IS *A10/10X*THE NUMBER OF Y-VA_U
       ZES FOR WHICH X≤GHAT =*I3/10X*THE IMMEDIATE X-VALUES BOUNDING GHAT
       ZARE*E10.4*,*E10.4/10X*THE ESTIMATE FOR THE INTERSECTION POINT,GHAT
       Z =*E10.4/10X*THE ESTIMATES OF ALPHA1,ALPHA2,BETA1,BETA2 ARE *3(E10
       Z.4*,*),E10.4/10X*ESTIMATE OF THE VARIANCE FOR THE SPLIT-LINE MODEL
       Z = *E10.4)
        RETURN
        END
```

D.

```
      SUBROUTINE UINTVAL(X,Y,NI,N,M,CSY)
      COMMON/A/XSUM(360),YSUM(360),XYSUM(360),X2SUM(360)
      COMMON/B/X1,X2,Y1,Y2,C1,C2
      COMMON/C/INTER,IA,IB
      COMMON/D/B1,B2,G
      COMMON/E/CT,DT,ET
      COMMON/F/J1,J2,I2
      DIMENSION X(1) , Y(1) , NI(1)
      INTEGER T

C     THE INTERVAL ESTIMATION PROCEDURE OF SECTION 3.7 .
C     THE INPUT VALUES X,Y,NI,N,M,CSY ARE DEFINED AS IN SUBROUTINE
C     HINKLEY

C     ROUTINES CALLED BEFORE ENTRY INTO UINTVAL - SORT , ACUM
C     ROUTINES CALLED WITHIN UINTVAL - WORK , UPDATE , FUNC
C     FUNCTIONS CALLED WITHIN UINTVAL - ZT

      CSX = X2SUM(M) - XSUM(M)**2 / N
C     STEP 1: FINDING AN INITIAL INTERVAL .
      L1 = M / 3 + 1
      L2 = M - L1

C     INTER : AN INDICATOR VALUE DEFINED IN SUBROUTINE UBOUND
      IF(INTER . EQ . 0) GO TO 10

      L1 = IA
      L2 = IB
   10 JJ1 = NI( L1 )
      JJ2 = NI( L2 )
      T =(L1 + L2) / 2
      CALL WORK(T,1,N,M,NI)
      M1 = NI(2)
      M2 = NI(M-1)
      IF(X(M1).GE.G.OR.G.GE.X(M2)) G = (X(JJ1) + X(JJ2)) / 2
      DIFF = AMIN1 ( G-X(1) , X(N)-G )/ 2
      A = G - DIFF
      B = G + DIFF
C     STEP 2: UPDATING THE INITIAL INTERVAL .
      CALL UPDATE (X,Y,NI,N,M,A,B)
C     CHECKING WHETHER THE UPDATED INTERVAL LIES WITHIN THE DATA RANGE
      IF(X(M1).LE.A.OR.B.LE.X(M2)) GO TO 20
      PRINT  90 , A , B ,X(J1) , X(J2)
      RETURN
   20 CONTINUE
      ZBOUND = 0
C     STEP 3: FINDING THE ESTIMATE OF GAMMA BY EMPLOYING METHOD 3.5 .
      I = 1
   30 I = I + 1
      JX1 = NI(I)
      JX2 = NI(I + 1)
      IF(X(JX2).LT.A) GO TO 30
```

```
      CALL WORK(I,3,N,M,NI)
      Z1 = ZT(G,G,CSX,B1,B2,CT,DT,ET)
      IF(Z1.LE.ZBOUND) GO TO 70
      IF(X(J1).GT.G.OR.X(I2).LT.G) GO TO 40
      GHAT = G
      T = I
      ZBOUND = Z1
      GO TO 70
   40 DF = DT - G*ET
      DG = (CT - G*DT) / DF
      IF(DF.GT.0.AND.DG.LT.X(I2).OR.DF.LT.0.AND.X(J1).LT.DG) GO TO 50
      GX = X(J1)
      IF(G.GT.X(I2)) GX = X(I2)
      ZX = ZT(GX,G,CSX,B1,B2,CT,DT,ET)
      IF(ZX.LE.ZBOUND) GO TO 70
      ZBOUND = ZX
      T = I
      GHAT = GX
      GO TO 70
   50 ZX1 = ZT(X(J1),G,CSX,B1,B2,CT,DT,ET)
      ZX2 = ZT(X(I2),G,CSX,B1,B2,CT,DT,ET)
      GX = X(J1)
      ZX = ZX1
      IF(ZX1-ZX2.GT.1.E-20) GO TO 60
      GX = X(I2)
      ZX = ZX2
   60 IF(ZBOUND.GT.ZX) GO TO 70
      ZBOUND = ZX
      T = I
      GHAT = GX
   70 IF(X(JX1).LT.B)GO TO 30
C     END OF SEARCHING PROCEDURE .
      CALL WORK(T,0,N,M,NI)
      XL = X(J1)
      XU = X(I2)
      CALL FUNC(GHAT,N,M,J1,T,RSS,AA1,BB1,BB2,CSY,1)
      AA2 = AA1 + GHAT*(BB1 - BB2)
      S = (CSY- (XYSUM(M)-XSUM(M)*YSUM(M)/N)**2/CSX - ZBOUND) / (N-4)
      CNAME = 10H  INTERVAL
      PRINT 80,CNAME,J1,XL,XU,GHAT,AA1,AA2,BB1,BB2,S
   80 FORMAT(1H1/10X*THE PROCEDURE USED IS *A10/10X*THE NUMBER OF Y-VA_U
     ZES FOR WHICH X≤GHAT =*I3/10X*THE IMMEDIATE X-VALUES BOUNDING GHAT
     ZARE*E10.4*,*E10.4/10X*THE ESTIMATE FOR THE INTERSECTION POINT,GHAT
     Z =*E10.4/10X*THE ESTIMATES OF ALPHA1,ALPHA2,BETA1,BETA2 ARE *3(E10
     Z.4*,*),E10.4/10X*ESTIMATE OF THE VARIANCE FOR THE SPLIT-LINE MODEL
     Z = *E10.4)
   90 FORMAT(/10X*SUBROUTINE ABORTED BECAUSE THE UPDATED INTERVA_*/1JX
     ZF10.4* - *F10.4*  LIES OUTSIDE THE RANGE OF THE DATA*F10.4* - *
     ZF10.4)
      RETURN
      END
```

E.

```
      SUBROUTINE UPDATE (X,Y,NI,N,M,A,B)
      DIMENSION X(1) , Y(1) , NI(1) , P(360) , C(5,5)
      COMMON/A/ XSUM(360),YSUM(360),XYSUM(360),X2SUM(360)

C     THIS ROUTINE PRODUCES THE UPDATED INTERVAL FOR USE IN ROUTINE
C     UINTVAL

C     ROUTINES TO BE CALLED BEFORE USING UPDATE- UINTVAL
C     ROUTINES CALLED WITHIN UPDATE - PROB

C     INPUT VALUES-
C      X:ARRAY OF THE N VALUES OF X .
C      Y:ARRAY OF THE N OBSERVATIONS .
C      C:THE MATRIX  (X≠X) .
C   N,M:AS IN THE OTHER ROUTINES .
C      A:INITIAL LOWER BOUND OF INTERVAL .
C      B:INITIAL UPPER BOUND OF INTERVAL .
C      P:ARRAY OF PROBABILITY VALUES FROM THE LOGISTIC DISTRIBUTION .
C     OUTPUT VALUES-
C      A:UPDATED LOWER BOUND OF NEW INTERVAL .
C      B:UPDATED UPPER BOUND OF NEW INTERVAL .

C     GHAT , CALCULATED BELOW ,IS AN UPDATED ESTIMATE FOR THE INTER-
C     SECTION POINT

C     CALCULATING THE ARRAY OF PROBABILITY VALUES
      E0 = (A + B) / 2
      E1 = (B - A) / 20
      CALL PROB (X,E0,E1,P,M,NI)

      I = 0
   10 I = I + 1
      J1 = NI(I)
      J2 = NI(I + 1)
      IF(X(J2).LE.A) GO TO 10
      IF(X(J1).LE.A.AND.A.LT.X(J2)) IL = I
      IF(X(J2).LE.B) GO TO 10
      IU = I + 1
      SP = NI(IL)
      SP2 = NI(IL)
      SPX = XSUM(IL)
      SP2X = XSUM(IL)
      SP2X2 = X2SUM(IL)
      SPX2 = X2SUM(IL)
      SPY = YSUM(IL)
      SPXY = XYSUM(IL)
      IL = IL + 1
```

```
      DO 20 I = IL,IU
      J2 = NI(I)
      J1 = NI(I) - NI(I - 1)
      SP = SP + J1*P(I)
      SPX = SPX + J1*P(I)*X(J2)
      SPX2 = SPX2 + J1*P(I)*X(J2)**2
      SP2 = SP2 + J1*P(I)**2
      SP2X = SP2X + J1*X(J2)*P(I)**2
      SP2X2 = SP2X2 + J1*(X(J2)*P(I))**2
      SPY = SPY + P(I)*(YSUM(I)-YSUM(I-1))
   20 SPXY = SPXY + P(I)*(XYSUM(I) - XYSUM(I - 1))
      C(1,1) = SP2
      C(1,2) = SP2X
      C(2,1) = SP2X
      C(2,2) = SP2X2
      C(1,3) = SP - SP2
      C(3,1) = SP - SP2
      C(1,4) = SPX - SP2X
      C(4,1) = SPX - SP2X
      C(2,3) = SPX - SP2X
      C(3,2) = SPX - SP2X
      C(2,4) = SPX2 - SP2X2
      C(4,2) = SPX2 - SP2X2
      C(3,3) = N - 2*SP + SP2
      C(3,4) = XSUM(M) - 2*SPX + SP2X
      C(4,3) = XSUM(M) - 2*SPX + SP2X
      C(4,4) = X2SUM(M) - 2*SPX2 + SP2X2
      C(1,5) = SPY
      C(2,5) = SPXY
      C(3,5) = YSUM(M) - SPY
      C(4,5) = XYSUM(M) - SPXY

C     USING THE IN-LINE SUBROUTINE FROM THE MATRIX PACKAGE WE FIND
C     THE INVERSE OF THE MATRIX ,C, AND AT THE SAME TIME WE CALCULATE
C     ESTIMATES FOR THE REGRESSION PARAMETERS, GIVING IN TURN THE
C     UPDATED ESTIMATE FOR THE INTERSECTION POINT
      CALL MATRIX (10,4,5,0,C,5,DET)

      GHAT = (C(1,5) - C(3,5)) / (C(4,5) - C(2,5))
      A = GHAT - 20*E1 / 3
      B = 2*GHAT - A
      RETURN
      END
```

```
      SUBROUTINE PROB(X,E0,E1,P,M,NI)
      DIMENSION X(1) , P(1) , NI(1)

C     ROUTINE TO BE CALLED BEFORE USING PROB - UPDATE

C     PROB GENERATES THE PROBABILITIES FOR THE INITIAL INTERVAL .
C     INPUT VALUES-
C      M:NUMBER OF DIFFERENT X VALUES .
C      X:ARRAY OF THE N DATA POINTS OF X .
C     E0:THE MEAN OF THE LOGISTIC DISTRIBUTION .
C     E1:STANDARD ERROR OF THE LOGISTIC DISTRIBUTION .
C     OUTPUT VALUES-
C     P:ARRAY OF PROBABILITY VALUES WHERE ,
C        P(I) = 0 FOR X(I) ≤ A (THE LOWER LIMIT OF THE INTERVAL) ,
C        P(I) = 1 FOR X(I) ≥ B (THE UPPER LIMIT OF THE INTERVAL ) .
C        P(I) = F(X) , WHERE F(X) IS THE CUMULATIVE LOGISTIC
C               DISTRIBUTION FUNCTION

      DO 10 I= 1,M
      J = NI(I)
      Z = -(X(J) - E0) / E1
      P(I) = 0
      IF(Z.LE.-10) GO TO 10
      P(I) = 1
      IF(Z.GE.10) GO TO 10
      P(I) = 1. - 1. / (1. + EXP(Z))
   10 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE  SORT( N,X,Y)
      DIMENSION  X(1) , Y(1)

C     THIS ROUTINE TAKES THE DATA , IN THE FORM ((X(I),Y(I),I=1,N) ,
C     AND SORTS IT IN ASCENDING ORDER, WITH RESPECT TO THE X-DATA

C     N : THE TOTAL NUMBER OF DATA POINTS
C     X : THE ARRAY OF X-DATA WHICH ON OUTPUT IS IN INCREASING ORDER
C     Y : ARRAY OF Y-DATA WHICH IS REORGANIZED IN ACCORD WITH YHE X-DATA

      N1 = N - 1
      DO 20 I = 1,N1
      I1 = I + 1
      K = I
      DO 10 J = I1,N
      IF(X(J).LT.X(K)) K = J
   10 CONTINUE
      T = X(I)
      X(I) = X(K)
      X(K) = T
      T = Y(I)
      Y(I) = Y(K)
   20 Y(K) = T
      RETURN
      END
```

```
      SUBROUTINE FUNC(U,N,M,J1,I,RU,AU1,BU1,BU2,CSY,IDENT)
      COMMON/A/XSUM(360),YSUM(360),XYSUM(360),X2SUM(360)
      COMMON/B/X1,X2,Y1,Y2,C1,C2
      REAL NK

C     IDENT IS AN INDICATOR VARIABLE WHERE ,
C     (1) FOR IDENT=0 WE CALCULATE THE RESIDUAL SUM OF SQUARES AS A
C        FUNCTION OF U-SEE EQUATION 3.2.(E)
C     (2) FOR IDENT=1 WE   ONLY  CALCULATE ESTIMATES OF THE
C        REGRESSION PARAMETERS FOR THE TWO LINES MEETING AT X = U. , FOR
C        THE FIXED PARTITION- I .

C     INPUT VALUE-
C         U:THE FIXED VALUE OF GAMMA.
C         N:NUMBER OF DATA POINTS.
C         M:NUMBER OF DIFFERENT X-VALUES , M≤N .
C        J1:THE NUMBER OF Y-DATA VALUES FOR WHICH X(I)≤U .
C         I:NUMBER OF DIFFERENT X VALUES FOR WHICH X(I)≤U.
C       CSY:CORRECTED TOTAL SUM OF SQUARES FOR THE Y-DATA

C     OUTPUT VALUES-
C      RU:THE RESIDUAL SUM OF SQUARES FOR GAMMA=U , CALCULATED ONLY IF
C        THE IDENTIFIER ( IDENT ) IS NOT EQUAL TO 1.
C     AU1:ESTIMATE FOR THE Y-INTERCEPT OF FIRST LINE.
C     BU1:ESTIMATE FOR SLOPE OF FIRST LINE.
C     BU2:ESTIMATE FOR SLOPE OF SECOND LINE.
C    NOTE;BOTH REGRESSION LINES MEET AT X=U →AU2=AU1+U (BU1-BU2) .

      J2 = N - J1
      NK = J1*FLOAT(J2)/N
      A = XYSUM(I)-J1*X1*Y1+NK*(Y1-Y2)*(X1-U)
      B = C1+NK*(X1-U)**2
      C = XYSUM(M)-XYSUM(1)-J2*X2*Y2-NK*(Y1-Y2)*(X2 - U)
      E = C1*C2+NK*(C2*(X1-U)**2+C1*(X2-U)**2)
      G = NK*(X1-U)*(U-X2)
      F = C2+NK*(X2-U)**2
      IF(IDENT.EQ.1) GO TO 10
      RU = CSY - (F*A**2 + B*C**2 - 2*C*G*A) / E
      RETURN
   10 BU1 = (F*A-G*C)/E
      BU2 = (B*C-G*A)/E
      AU1 = Y1-BU1*X1 + J2*( (Y2-Y1) + BU1*(X1-U) - BU2*(X2-U)) / N
      RETURN
      END



      FUNCTION ZT(U,GHAT,CSX,B1,B2,CT,DT,ET)

C     THIS FUNCTION EVALUATES EQUATION 3.3.A(5).

C     INPUT  VALUES-
C            U:THE FIXED VALUE FOR GAMMA.
C         GHAT:THE ESTIMATE FOR THE INTERSECTION POINT OF THE
C              TWO REGRESSION LINES
C          CSX:CORRECTED TOTAL SUM OF SQUARES FOR ALL THE X-DATA.
C        B1,B2:UNCONSTRAINED LEAST SQUARES ESTIMATES FOR THE SLOPES
C              OF THE TWO STRAIGHT LINES.
C     CT,DT,ET: ARE DEFINED IN EQUATION 3.2.(E)


      ZT = (B1-B2)**2/(CT-2*DT*U+ET*U**2)*(CT-DT*(U+GHAT)+ET*U*GHAT)**2
     Z     /CSX

      RETURN
      END
```

```
      SUBROUTINE ACUM(N,M,X,Y,NI,CSY)
      COMMON/A/XSUM(360),YSUM(360),XYSUM(360),X2SUM(360)
      DIMENSION X(1),Y(1),NI(1)

C     THIS ROUTINE CALCULATES ACCUMULATING SUMS,SUMS OF PRODUCTS AND
C     SUMS OF SQUARES OF THE (X,Y) DATA - AFTER THE DATA HAS BEEN
C     THROUGH THE SORT ROUTINE .

C     INPUT VALUES-
C            N:NUMBER OF DATA POINTS .
C            M:NUMBER OF DIFFERENT X VALUES , M≤N .
C            X:ARRAY OF ORDERED X VALUES .
C            Y:CORRESPONDING ARRAY OF Y VALUES .

C     OUTPUT VALUES -
C          NI:AN ARRAY , WHERE THE ELEMENTS NI(I) REPRESENT THE
C             NUMBER OF OBSERVATIONS OF Y-DATA FOR WHICH THE
C             CORRESPONDING X-VALUES ARE ≤ TO X(I)
C         CSY:CORRECTED SUM OF SQUARES FOR THE TOTAL Y-DATA
C      XSUM(I):=NI(1).X(1)+.............+(NI(I)-NI(I-1)).X(I)
C      YSUM(I):=Y(1)+Y(2)+.....+Y(K) ,K=NI(I)
C     XYSUM(I):=X(1).(Y(1)+Y(2)+.......Y(NI(1)))
C               +X(2).(Y(NI(1)+1)+.....+Y(NI(2)))+.....
C               ...+X(I).(Y(NI(I-1)+1)+.......+Y(NI(I))) .
C     X2SUM(I):=NI(1).X(1)**2+.........+(NI(I)-NI(I-1)).X(I)**2  .
      DO 10 I = 2,N
   10 NI(I)=0
      XSUM(1) = X(1)
      YSUM(1) = Y(1)
      XYSUM(1) = X(1)*Y(1)
      X2SUM(1) = X(1)**2
      SYY = Y(1)**2
      NI(1) = 1
      I = 1
      DO 20 L=2,N
      IF(ABS(X(L)-X(L-1)).GE.1.E-6) I = I+1
      IF (I.NE.1) GO TO 20
      SYY = SYY + Y(L)**2
      X2SUM(1) = X2SUM(1) + X(L)**2
      XYSUM(1) = XYSUM(1) + X(L)*Y(L)
      YSUM(1) = YSUM(1) + Y(L)
      XSUM(1) = XSUM(1) + X(L)
   20 NI(I) = NI(I) + 1
      M = I
      DO 30 I = 2,M
   30 NI(I) = NI(I-1) + NI(I)
      DO 50 I = 2,M
      J1 = NI(I-1) + 1
      J2 = NI(I)
      SX= 0
      SY = 0
      SXY = 0
      SX2 = 0
      DO 40 JK = J1,J2
      SX = SX + X(JK)
      SY = SY + Y(JK)
      SXY = SXY + X(JK)*Y(JK)
      SX2 = SX2 + X(JK)**2
   40 SYY = SYY + Y(JK)**2
      XSUM(I) = XSUM(I - 1 ) + SX
      YSUM(I) = YSUM(I - 1) + SY
      XYSUM(I) = XYSUM(I - 1) + SXY
      X2SUM(I) = X2SUM(I - 1) + SX2
   50 CONTINUE
      CSY = SYY - YSUM(M)**2 / N
      RETURN
      END
```

```
      SUBROUTINE WORK(I,IPAR,N,M,NI)
      COMMON/A/XSUM(360),YSUM(360),XYSUM(360),X2SUM(360)
      COMMON/B/X1,X2,Y1,Y2,C1,C2
      COMMON/D/B1,B2,G
      COMMON/E/CT,DT,ET
      COMMON/F/J1,J2,I2
      COMMON/H/A1,A2
      DIMENSION NI(1)

C     THIS ROUTINE CALCULATES VARIOUS VARIABLES FOR A FIXED PARTITION,I.
C     THE FIRST PARTITION BEING , (X(1),Y(1)) ,,,,,,,,(X(J1),Y(J1))  .
C     THE SECOND PARTITION BEING , (X(J1+1),Y(J1+1)),,,,(X(N),Y(N))  .
C     J1:TOTAL NUMBER OF OBSERVATIONS OF THE FIRST PARTITION .
      J1 = NI(I)
C     J2:NUMBER OF OBSERVATION OF THE SECOND PARTITION .
      J2 = N-J1
C     I2:NUMBER OF X-VALUES FOR WHICH X≤ X(NI(I+1))
      I2 = NI(I+1)
C     X1:THE MEAN OF THE X DATA FOR THE FIRST PARTITION .
      X1 = XSUM(I)/J1
C     X2:MEAN OF X DATA FOR THE SECOND PARTITION .
      X2 = (XSUM(M)-XSUM(I))/J2
C     Y1:MEAN OF Y DATA FOR FIRST PARTITION .
      Y1 = YSUM(I)/J1
C     Y2:MEAN OF THE Y-DATA FOR THE SECOND PARTITION
      Y2 = (YSUM(M)-YSUM(I))/J2
C     C1:SUM OF SQUARES FOR X DATA OF FIRST PARTITION .
      C1 = X2SUM(I)-J1*X1**2
C     C2:SUM OF SQUARES FOR X DATA OF SECOND PARTITION .
      C2 = X2SUM(M)-X2SUM(I)-J2*X2**2

      IF(IPAR.EQ.0) RETURN
C     B1:ESTIMATE OF SLOPE FOR FIRST PARTITION .
      B1 = (XYSUM(I)-J1*X1*Y1)/C1
C     A1:ESTIMATE FOR Y-INTERCEPT OF FIRST PARTITION .
      A1 = Y1 - B1*X1
C     B2:ESTIMATE OF SLOPE FOR SECOND PARTITION .
      B2 = (XYSUM(M)-XYSUM(I)-J2*X2*Y2)/C2
C     A2:ESTIMATE FOR Y-INTERCEPT OF SECOND PARTITION .
      A2 = Y2 - B2*X2
C      G:ESTIMATE FOR THE INTERSECTION POINT ,GAMMA,FOR THIS PARTITION .
      G = (A1-A2) / (B2-B1)

      IF(IPAR.EQ.1) RETURN
      W = J1*FLOAT(J2) / N
      ET = W*(C1 + C2)
      DT = W*(X1*C2 + C1*X2)
      CT = C1*C2 + W*(C1*X2**2 + C2*X1**2)
      RETURN
      END
```

BIBLIOGRAPHY

1. BACON, D.W. and WATTS, D.G. (1971).
   Estimating the transition between two intersecting straight lines.
   Biometrika, 58, 3, p.525.

2. BHATTACHARYYA, G.K. and JOHNSON, R.A. (1968).
   Non-parametric tests for shift at an unknown time point.
   Ann. Math. Statist. 39, p. 173-43.

3. BROWN, R.L. and DURBIN, J. (1968).
   Methods of investigating whether a regression relationship is constant over time. European Meeting, 1968. Selected Statistical Papers, I, pp.37-45.
   Amsterdam: Mathematisch Centrum.

4. DEMPSTER, A.P.
   Elements of Continuous Multi-variate analysis.
   Addison-Wesley Pub. Co.

5. DOBBING, J.
   Scientific Foundations of Paediatrics, p.565-76.
   W. Heinemann Medical Books Ltd., London.

6. FEDER, D.J. and SYLVESTER, D.L. (1968).
   On the asymptotic theory of Least Squares estimation in segmented regression: identified case (Abstract).
   Ann. Maths. Statist., 39, p.1362.

7. HINKLEY, D.V. (1969).
   Inference about the intersection in two-phase regression.
   Biometrika, 56, 3, p.495.

8. HUDSON, D.J. (1966).

Fitting segmented curves whose join points have to be estimated.

J. Am. Statist. Ass., 61, p.1097-129.

9. KALMAN, R.E. (1960).

A new approach to linear filtering and prediction problems.

Trans. Am. Mech. Eng., J. Basic Engineering, 82, p.35-45.

10. McGEE, V.E. and CARLETON, W.T. (1970).

Piecewise regression.

J. Am. Statist. Ass., 65, p.1109-21.

11. PAGE, E.S. (1955).

A test for a change in a parameter occurring at an unknown point.

Biometrika, 42, p.523-7.

12. QUANDT, R.E. (1960).

Tests of the hypothesis that a linear regression system obeying two separate regimes.

J. Am. Statist. Ass., 55, p.324-30.

13. RAO, C.R.

Linear Statistical Inference and its Applications; 2nd edition.

Pub. J. Wiley and Sons.

14. WAINER, H. (1971).

Piecewise Regression. A Simplified Procedure.

Br. J. Math. Statist. Psychol., 24, p.83-92.