FITTING SEGMENTED LINEAR-REGRESSION MODELS

A Dissertation submitted to the University of Kent at Canterbury as part fulfilment of the award of the degree of

M.Sc. in Statistics

by

S.A.S. Olorunju

April 1981

ACKNOWLEDGEMENTS

I wish to acknowledge my Supervisor, Mr J.M. Bremner, for his co-operation and assistance in the course of writing and the organisation of the Dissertation. I gratefully acknowledge Miss Janet Harris whose words of encouragement and company I cherished very much during the period of my studies here.

Finally, I express my gratitude to my Sponsors - National Animal Production Research Institute (NAPRI) Zaria, Nigeria, for financial support and, in particular, Dr. Mike Olayiwole for his words of encouragement. Also to Mrs Eve Hurste for a job beautifully done.

CONTENTS

1.	INTRODUCTION	1
2.	DEFINITIONS AND DESCRIPTIONS OF SEGMENTED REGRESSION PROBLEM	3
2.1	Definitions of Relevant Terminologies	3
2.2	Statement of the Problem	3
2.3	Applications	10
3	ALGORITHM FOR ESTIMATION OF PARAMETERS IN THE 2-SEGMENT REGRESSION MODEL	13
3.1	Brief Introduction	13
3.2	Developments and Definitions of Basic Concepts	14
3.2.1	General Discussion of Algorithms	17
3.3	Theorem for 2-Segment (Linear) Regression Problem	18
3.3.1	Description of Hudson's Algorithm	20
3.3.2	Description of Hinkley's Algorithm	21
3.4	The Improved Algorithm	22
3.5	Illustration of the Use of the Improved Algorithm	24
3.6	Further Improvement and General Discussion on the Algorithms	33
4	OTHER ESTIMATION TECHNIQUES FOR CONTINUOUS MODELS	37
4.1	General Discussion	37
4.2	Graphical Technique (Plotting)	38
4.3	Grid Search	39
4.4	Bayesian Technique	40
5	ESTIMATION ALGORITHMS FOR PIECEWISE MODELS	45
5.1	General Description	45
5.2	Estimation Technique for 2-Segment (equal segment variance) Piecewise Regression Model	46
5.3	Extension to K>2 Segments (equal segment variance)	47
5.4	Estimation Technique for 2-Segment (unequal segment variance) Piecewise regression models	52

5.5 Other Estimation Techniques for K>2 Segment Piecewise Regression Models 53

5.5.1 The Search Algorithms 53

5.5.2 Hierarchical Clustering Technique 54

5.6 Selection of the Number of Segments in Piecewise Regression Models 58

6. INFERENCES IN SEGMENTED AND PIECEWISE REGRESSION MODELS 62

7. SUMMARY AND CONCLUSION 67

REFERENCES 69

APPENDIX 71

INTRODUCTION

In regression analysis problems (studies of relationships) the expectations of a response variable Y is generally represented by a polynomial function of the independent (regressor) variable x, over the entire range of data being analysed. However, such representations sometimes prove inadequate because of changes which occur within the system. Often, these changes cannot be accounted for or explained explicitly.

Some of these situations have been dealt with more adequately by fitting a model, consisting of several submodels, in place of a single polynomial. Each submodel is only applicable over a limited range of values of the independent variable. The points where one submodel succeeds another one are called change-points. The locations of these points are frequently unknown, in which case it is necessary to obtain the estimates of these points/range in addition to the other parameters of each submodel present in the overall model.

This Dissertation looks at diverse approaches that have been used in solving problems of this type with only one single regressor variable. Consideration is given only to linear-regression submodel problems.

We shall consider two situations, the first where the overall model is continuous and the second when it is discontinuous. The overall model is said to be continuous whenever the expectation of the response variable Y is a continuous function of the independent variable x.

The term "join-point" will be used for the change-point in the content of continuous overall models. Similarly, "jump-point" will be used for change-points in the context of discontinuous models.

Chapter 2 looks at the definitions of the K-segmented or Piecewise regression problem as well as the introduction, developments of the ideas and terminologies for use in the Dissertation. Chapter 3 considers the specific problems related to the case of a continuous model with only two

linear segments. At this stage various algorithms that have been proposed are reviewed. In addition, an improved algorithm is proposed which like the others is based upon the theory of standard least squares. The use of the improved algorithm is illustrated by its application to data from some practical experiments and some artificially constructed examples.

Chapter 4 looks at other estimation procedures for continuous models. Chapter 5 proceeds to look at the estimation approaches for discontinuous model problems - Piecewise regression. Finally Chapter 6 discusses inferences in general for these types of regression problem. For the purpose of this Dissertation, definitions have been given for most of the terms adopted for a clear understanding of what the study is all about because quite a number of the terms have been observed used interchangeably in the literature survey of work undertaken in this area.

2. DEFINITIONS AND DESCRIPTION OF SEGMENTED REGRESSION PROBLEM

2.1 Definitions of relevant terminologies

In the literature, several terms have been adopted and used interchangeably. However, it is proposed here to adopt terms having specific definitions in order to avoid confusion. There are two basic terms that will occur frequently in the Dissertation. The terms are "segmented" and "Piecewise".

"Segmented" is used for regression models with continuous model, while the term "Piecewise" is used for regression models which are discontinuous. Hence, as explained in the previous chapter, the term "joinpoint" denotes the change-point in segmented regression models. Much of this Dissertation considers segmented regression models because they are most appropriate to the majority of practical applications.

Apart from the above definitions, the term "Partition" will also be used frequently to denote a division of a set of n data points $(x_1, y_1).....(x_n, y_n)$ into disjoint subjects of adjacent points. The term is used both in the context of Segmented and Piecewise regression models.

2.2 Statement of the problem

The problem of a K-segmented regression is defined in terms of linear segments only. This is because most consideration is given to it in this Dissertation. Though some authors have considered models with polynomial segments, few applications of such models have been made.

Suppose there are n observations represented by $(x_1, y_1)....(x_n, y_n)$ where $x_1<x_2<...<x_n$ to which the relationship between the response variable Y and the independent variable x is given by

$$
\begin{aligned}
E(Y) &= f_1(x) = \beta_{11} + \beta_{12}x \qquad -\infty<x<\gamma_1 \\
&= f_2(x) = \beta_{21} + \beta_{22}x \qquad \gamma_1<x<\gamma_2 \qquad (2.2.1) \\
&= - - - - - - - - - \\
&= f_k(x) = \beta_{k1} + \beta_{k2}x \qquad \gamma_{k-1}<x<\infty
\end{aligned}
$$

where $f_j(x)$ $(j = 1...k)$ denotes the j^{th} segment and $\gamma_i(j = 1...k-1)$ denotes the j^{th} join-point (which is not necessarily known); the x_i's are distinct. To simplify some of the following theory, it will be assumed that x's are distinct, though simple extension will enable us to deal with repeated x-values. Model (2.2.1) represents a general K-linear segment regression model. In the context of segmented regression analysis, model (2.2.1) is continuous and subject to the constraint

$$f_j(Y_j, \underline{\beta}_j) = f_{j+1}(Y_j; \underline{\beta}_{j+1}) \quad j = 1...k-1 \qquad (2.2.2)$$

Given n observations to which model (2.2.1) is to be fitted, the Statistician is faced with one immediate problem in that γ_j's whose values are not necessarily known is involved. When γ_j's are unknown, (2.2.1) is non-linear in γ_j. Consider a 2-linear segment example given below for an illustration.

EXAMPLE

Suppose we fit the following 2-linear segment model to the data:

$$
\begin{aligned}
E(Y) &= f_1(x) = \beta_{11} + \beta_{12}x \qquad x<\gamma \\
&= f_2(x) = \beta_{21} + \beta_{22}x \qquad x\geq\gamma \qquad (2.2.3)
\end{aligned}
$$

subject to the constraint

$$\beta_{11} + \beta_{12}\gamma = \beta_{21} + \beta_{22}\gamma \qquad (2.2.4)$$

Model (2.2.3) is non-linear in γ if its value is unknown and (2.2.4) does not involve linear function of the parameters either. Hence the problem is more involved than the simple linear regression analysis.

Suppose we look closely at the two linear segment example defined above for the n-observations. The problem of non-linearity may be

appreciated if we look at the behaviour of the function defined by

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where y_i's are the fitted values of the response variable at point x_i. We shall refer to S as the "sum of squares function" which is a function of all the parameters. Linear least-squares method enables us to minimize S with respect to all parameters but γ. But γ is a function of the parameters. Therefore, minimizing S with respect to the parameters is equivalent to minimizing the function RSS(γ) which is the residual sum of squares expressed as a function of the point γ for the model defined in (2.2.3) above. For the two-linear segment problem, it is observed that a graph of RSS(γ) against γ does not result in a function which possesses a continuous derivative at every point on the curve that arises in a linear-regression problem. It has several gradient discontinuities at the observed adscissae values which result in a series of joined curves. This is because the RSS (γ) is not differentiable with respect to γ at the x-values- where the gradient discontinuities occur.

For illustration, the two graphs relate to a two-linear segment example. Fig. 2.2.1a is the plot of data given in the example. Fig. 2.2.1b gives a plot of RSS(γ) against all values of γ, $x_1<\gamma<x_n$. The details of the expression for the function as well as the calculation are discussed in Section 3.2. The interesting point is the visual display of the problem of non-linearity associated with regression problem when γ_j's are not known.

Now, from Fig. 2.2.1a, the plot of the data seems to suggest a more or less perfect two-linear relationship (in fact, as will be seen later, the parameter values give an exact fit). As indicated earlier, the function does not possess continuous derivative at every point on the graph. Furthermore, notice the constant value of RSS(γ) for $x_1 \leq \gamma \leq x_2$, as

Example 1.1

X	2.0	3.0	4.0	5.0	7.0	8.0	10.0	11.0	12.0	14.0	16.0
Y	16.0	14.0	12.0	10.0	6.0	4.0	8.0	10.0	12.0	16.0	20.0

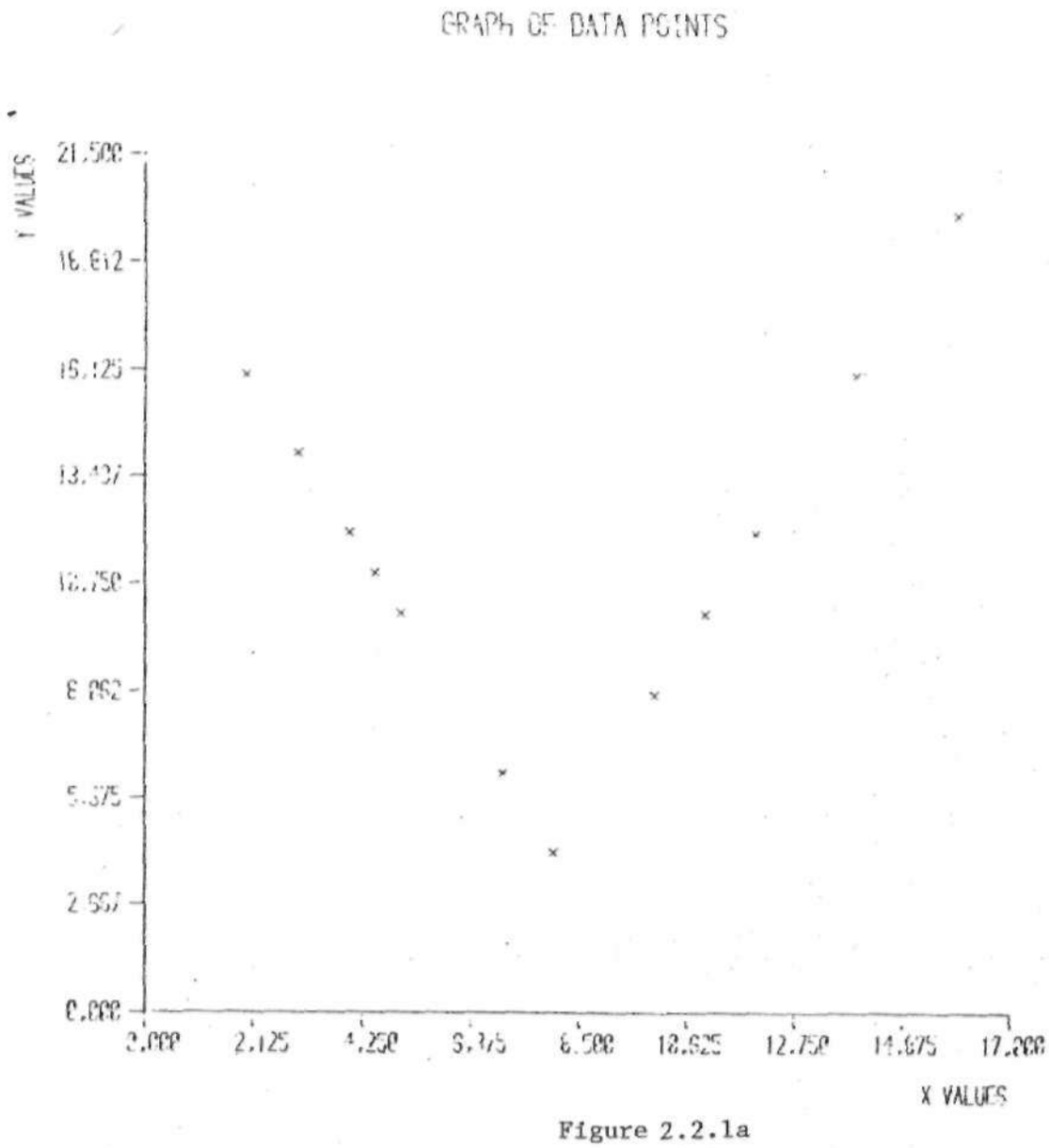


Figure 2.2.1a

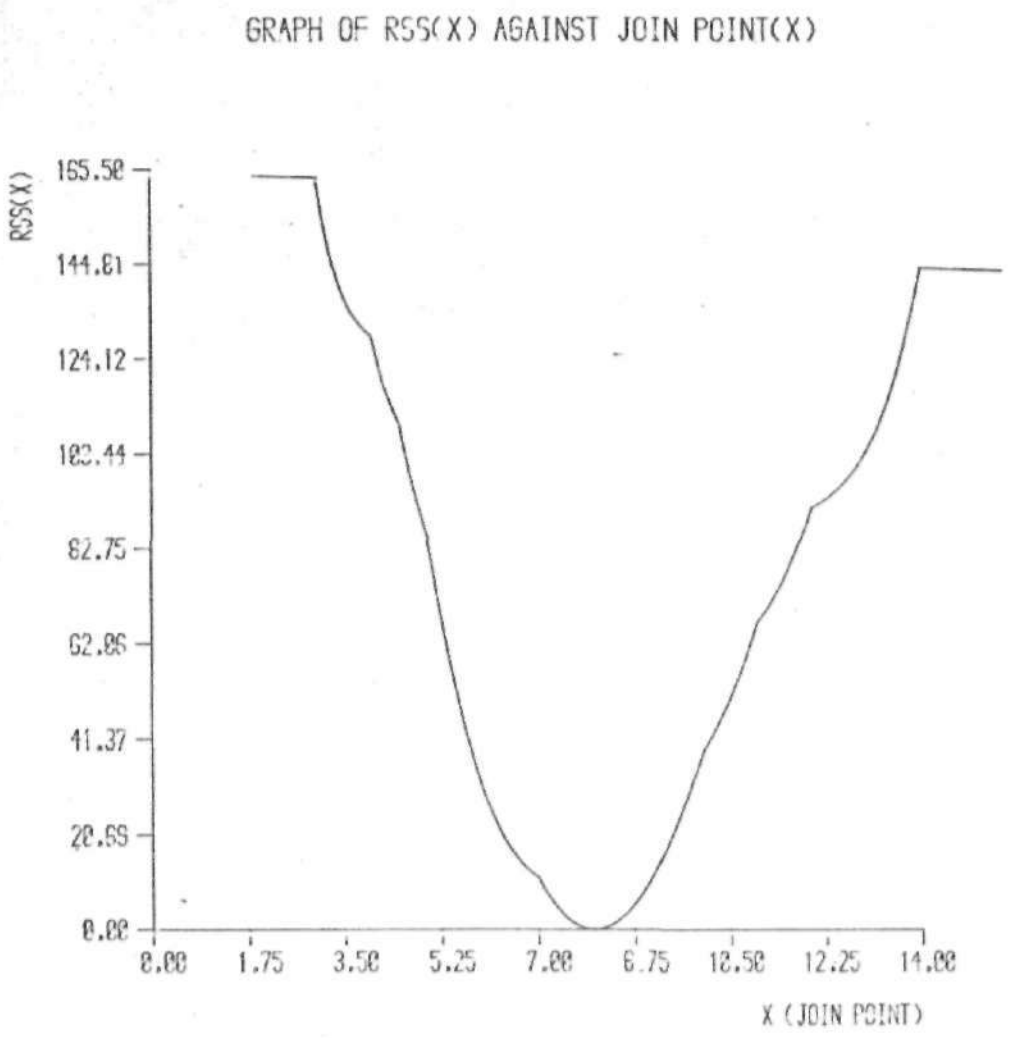

GRAPH OF RSS(X) AGAINST JOIN POINT(X)
RSS(X)
165.50
144.81
124.12
103.44
82.75
62.06
41.37
20.69
0.00
0.00
1.75
3.50
5.25
7.00
8.75
10.50
12.25
14.00
X (JOIN POINT)

Figure 2.2.1b

well as $x_{n-1} \leq \gamma \leq x_n$. If we consider the first region given by the interval $[x_1, x_2]$, the constant behaviour of RSS(γ) is explained by the fact that for values of γ within the specified range above, there is only one data point constituting a subset to which the first segment is fitted - (x_1, y_1), while the second line is fitted to the remaining points, implying that the value of the residual sum of squares is strictly determined by the second segment containing points $\{(x_2, y_2) \ldots . (x_n, y_n)\}$. The residual sum of squares corresponding to the left hand segment with only one point is zero since there is always a perfect fit there. A similar argument holds for the interval $[x_{n-1}, x_n]$. Essentially, the flat part of the graph does not contribute to the determination of the location of the minimum of the function RSS(γ). Furthermore, no unique solution is obtained for the slope of the fitted line for the segments in these two ranges.

In order to avoid this non-uniqueness, we shall impose a restriction in fitting linear segment regression models so that each segment is fitted to at least two data points. Hence no more attention will be given to the regions $[x_1, x_2)$ and $(x_{n-1}, x_n]$.

Now consider the regions between $x_2 < \gamma < x_{n-2}$. It is noticed that the incidence of gradient discontinuities occurs over all the points of the observed x-values. Between any adjacent observed x-values, the function RSS(γ) possesses a continuous derivative. The value of RSS(γ) decreases as γ increases. It attains its minimum at $\gamma = 8$ with an RSS(γ) = 0, then starts to rise again to attain another maximum at $\gamma = 14$. It is not common in practice to have the residual sum of squares exactly zero. However, it is not surprising here, as indicated earlier, because of the exact fit of fitted model to the data (Fig. 2.2.1a). Recall that this example was constructed artificially to illustrate the non-linearity problem defined above by considering the behaviour of the residual sum of squares functions for various values of γ for a two-linear segment

example. Notice also that the function possesses a continuous derivative within the region in which the minimum of RSS(γ) is attained in the above graph Fig. 2.2.1b (i.e. between 7 and 8.75 on graph). This will be found not to be the case for every problem. For this particular example, it is necessary to point out the fact that RSS(γ) is differentiable at $\gamma = 8$ (which is one of the observed x-values). This example is therefore atypical.

In the two-segment examples to be looked at in Chapter 3, consideration will only be given to the behaviour of the function RSS(γ) for all values of γ satisfying $x_2 \leq \gamma \leq x_{n-2}$. It might be worth emphasizing the fact that in situations where the values of γ_j's are known, the problem is no longer non-linear and hence ordinary least sequences theory is applicable. Fuller (1969) followed such an approach when he fitted the trend line to some adjusted corn yield data. Looking at the graph of RSS(γ) against all possible trial values of γ, we shall be looking at this approach in detail later in Chapter 4. Alternative approaches to the solution of the continuous problem are considered there as well as in Chapter 3.

The problem defined by the model (2.2.1) can be looked at on the following broad bases. Firstly, it is essential to know whether or not the overall model is assumed to be continuous. Also one wishes to know whether or not the number of segments is known or at least can be obtained from, for example, such preliminary techniques as plotting the data. Thirdly, one will be interested in knowing whether or not the join-points are known. Usually, there are no such assumptions about the prior knowledge of the join-points. In fact, one of the aims of some experiments applying the technique of segmented regression models is to estimate these points. Finally, one has to make the decision on whether the submodels should have equal or distinct variances.

In the light of the points made above, several techniques have been developed. Each of the techniques made its own assumption regarding the points discussed above. One general point concerning most of the techniques for continuous models is that they assume the number of segments is known. For the major part of this Dissertation and in particular for the segmented regression model, it is assumed that the number of segments can be estimated from the plot of the data if it is not already known.

2.3 Applications

Segmented and Piecewise regressions have a broad range of applications in many spheres of research work. These range through applied statistics in general from agricultural studies to industrial and medical research. In this section, we consider some examples of areas where the techniques have been used. It is necessary to point out at this stage that most of the examples found in the literature survey gave no theoretical reason or argument for applying segmented or Piecewise regression models instead of polynomial regression models. Most users decided on their use of segmented regression models from a plot of the data collected.

One important area in which segmented regression models have been used is in growth studies of plants as well as growth studies of animals when they are subjected to specified treatments, such as the application of nutrients over a period of time. These methods have been used in analysing the data collected in poultry experiments. Segmented regression models are particularly useful in growth studies.

Generally, in growth studies, it has been observed that the growth in plants or animals is not consistent over the life of the plants or animals. Because of lack of consistency, it is considered more appropriate and reasonable that each stage of the growth be modelled independently using segmented regression models. So by considering each stage of growth

separately, the overall growth can then be adequately modelled by some linear relationships which are easier to deal with than polynomials.

Sprent (1961) fitted a segmented regression model to data obtained from counts of the number of nodes in terminal buds of spurs of a variety of apples. Using the segmented regression modelling approach, Sprent remarked that biologists often prefer models in which, for example, each segment is a linear regression, rather than there being one polynomial function. No theoretical reason was given for the preference apart from the ease in mathematical manipulations.

Fuller (1969) demonstrated the use of segmented regression techniques in the analysis of a fertilizer experiment where the objective was to estimate the response surface for corn yield as a function of nitrogen and phosphorus. This is an example with two independent variables which lies outside the scope of this Dissertation and will not be considered any further.

Fisher et al. (1973) applied the theory of segmented regression to biological data in which the response of laying hens (in terms of number of eggs laid) was related to their intake of aminoacids (the independent variable).

In an unpublished paper of the National Animal Production Research Institute (Nigeria), the ideas of segmented regression procedure were used to model the rate of growth of day-old chickens over a period of six months. The objective of the experiment was to choose the composition of nutrients which favours the highest rate of growth of the group of birds in cages. The study ascertained the effect of the gradual change of the composition of the nutrient in the feeds used for the chickens. For various compositions the corresponding weights were recorded. On plotting the data, segmented regression models were seen to be appropriate in order to model the growth. Other biological applications include Gallant and

Fuller (1973) who modelled agricultural data by segmented regression technique.

Amongst industrial applications, Beckman and Cook (1979) used a segmented regression model to describe the relationship between the yield of hydrogen sulphide from gamma radiolysis of ethyl mercaptan in mixtures of some inert gases (neon, argon and krypton) to krypton pressure (the independent variable). Their data is considered further in Section 3.5. Smith (1979) also gave an industrial application on the study of conductance obtained from a chemical conductivity experiment in which a strong monotropic acid (hydrochloric acid) was titrated against a known concentration of base. The conductivity was taken as the response variable and the concentration as the independents. The objective was to determine the equivalence point which, in the context defined by model (2.2.3) is the join-point. This implies that the problem essentially requires two-linear segments, where the first is fitted to the basic medium and the second when the solution becomes acidic; that is when the equivalence (neutral) point is exceeded.

For social studies, Eppright et al. (1972) considered the application of segmented regression procedures to study the relationship between the weight-to-height ratio of pre-school boys and their age. The objective of the studies was to obtain information about the nutrition requirements of infants and pre-school children.

An application of Piecewise regression models was made by McGee and Carleton (1970). They analysed some financial data based on the dollar volume of the states of the Boston Stock Exchange. Hawkins (1976) used the same set of data for the validation of his proposed algorithm (to be discussed in Chapter 5).

3. ALGORITHMS FOR THE ESTIMATION OF PARAMETERS IN THE 2-LINEAR SEGMENT REGRESSION MODEL

3.1 Introduction

The estimation of the parameters in a continuous regression model which consists of K segment as defined in (2.2.1) has not been dealt with in much depth by many investigators. Most familiar approaches are based on the standard linear least squares methods. Pioneering work has been done by Hudson (1966). He introduced the preliminary ideas on how to solve the problem for the more general use of several segments, in which the segments are polynomials and the join-points are unknown. Furthermore, he proposed an algorithm for two-segment problems.

In this Chapter, consideration will be given to the development of ideas based on two segments alone. Some basic definitions to be used later in the Chapter are introduced along with some other basic concepts used in this Dissertation.

Also, we shall look at Hudson's (1966) and Hinkley's (1969, 1971) algorithms, developed for solving the two-segment problem. The algorithms apply estimation techniques based on linear least squares theory. An improved algorithm is proposed for solving two-segment problems. Its operation and validation are illustrated by applying it to some constructed and practical examples. While Hudson and Hinkley considered the situation in which the segment variances are equal, Singpurwalla (1974) used Hudson's ideas for a two-segment model in which the error variance is a function of the mean. However, his approach is not based on the standard linear least squares theory as Hudson and Hinkley. The example used to illustrate the operation of the improved algorithm has an underlying assumption that segment variances are equal, like Hudson and Hinkley.

3.2 Development and Definitions of Basic Concepts

Suppose there are n data points whose observations are given by $(x_1, y_1).....(x_n, y_n)$ with the x's ordered and distinct. Suppose further that two linear segments are to be fitted to the data. Let the first line be fitted to the points $(x_1, y_1).....(x_i, y_i)$ and the second line to the remaining points $(x_{i+1}, y_{i+1}).....(x_n, y_n)$. In this Dissertation a division of the n data points in disjoint subsets will be referred to as a partition; the partition above will be referred to as partition i. The partition to be considered will correspond to values of i, satisfying the condition $2 \leq i \leq n - 2$; this condition implies that each segment will be fitted to at least two points, and is imposed for the reasons given in Section 2.2. Now, let $f_j(x; \underline{\beta}_j)$ $(j = 1,2)$ denote the expectation of $\underline{Y}$ given by the submodel for the j^{th} segment, $\underline{\beta}_j$ being the corresponding vector of parameters $\underline{\beta}_j = (\beta_{j1}, \beta_{j2})$, $\underline{\beta}'_1 = (\beta_{11}, \beta_{12})$ and $\underline{\beta}'_2 = (\beta_{21}, \beta_{22})$. Let γ be the parameter for the join-point and RSS to denote the residual sum of squares obtained for any such partition under consideration.

Local Least Squares Estimation This term is used to denote the procedure by which the estimates of the parameters for partition i are obtained, by fitting models separately to each disjoint subset of data points. The estimates of parameters for partition i are called local least squares estimates. As a vector, these will be denoted by $\underline{\beta}_{LLs}(i)$. It is to be noted that $\hat{\underline{\beta}}_{LLs}(i)$ represents the estimates of the parameter for the two segments. This implies that $\hat{\underline{\beta}}_{LLs}(i)$ contains four of this section. The models are fitted separately to the subsets as if the data were not from the same experiment; that is, each subset if a "piece". In local least squares estimation, the estimates of the join-point denoted $\gamma_{LLs}(i)$ for partition i is given by the abscissae of the intersection of the fitted local least squares lines obtained by solving the equation:

$$\hat{\beta}_{11} + \hat{\beta}_{12}\gamma = \hat{\beta}_{21} + \hat{\beta}_{22}\gamma$$

where $\hat{\beta}_{11}$, $\hat{\beta}_{12}$, $\hat{\beta}_{21}$ and $\hat{\beta}_{22}$ are the appropriate local least squares estimates for the parameters of the two lines. Note that $\hat{\gamma}_{LLs}(i)$ need not lie in the range $[x_i, x_{i+1}]$ for partition i.

Constrained least squares estimation

The estimation is performed by least squares procedures, but the fitted lines are constrained to intersect at some predetermined x-value. In this case the type of estimation required is rather more involved than that for the local least squares estimation. The estimation procedure is developed below for the two-linear segment case. The more general linear model for the full rank case is given in the Appendix.

Construction of constrained least squares estimates for two-linear segments

Suppose we are required to fit two linear segment models to the data $(x_1, y_1)....(x_n, y_n)$. Suppose we fit the model

$$f(x) = \beta_{11} + \beta_{12}x \text{ for observations } 1, 2, ... i$$

$$= \beta_{21} + \beta_{22}x \text{ for observations } (i+1) ..... n,$$

subject to the constraint that the lines corresponding to these two submodels intersect at $x = \gamma$. Note that, although we shall generally apply the result with $\gamma \in [x_i, x_{i+1}]$, no such restriction is imposed in what follows. If $\underline{\beta}_1$ and $\underline{\beta}_2$ denote the parameters of the two submodels, the local (unconstrained) least squares estimates of $\hat{\underline{\beta}}_1$ and $\hat{\underline{\beta}}_2$ are

$$\hat{\underline{\beta}}_j = C_j^{-1}\underline{X}'_j\underline{Y}_j \quad (j = 1,2) \text{ where } \underline{C}_j = \underline{X}'_j\underline{X}_j,$$

$\underline{X}_j$ is the appropriate design matrix for the first and second subset of the data and $\underline{Y}_j$ the appropriate vector of responses. The residual sum of squares for the local least squares fit is obtained by adding the two individual residual sums of squares for the data subsets that is

$$RSS = \|\underline{Y}_1 - \underline{X}_1\hat{\underline{\beta}}_1(i)\|^2 + \|\underline{Y}_2 - \underline{X}_2\hat{\underline{\beta}}_2(i)\|^2 .$$

Now consider the constraint that the two lines intersect at γ, i.e.

$$\beta_{11} + \beta_{12}\gamma = \beta_{21} + \beta_{22}\gamma$$

This constraint is linear in $\underline{\beta}_1 = \begin{bmatrix} \beta_{11} \\ \beta_{12} \end{bmatrix}$ and $\underline{\beta}_2 = \begin{bmatrix} \beta_{21} \\ \beta_{22} \end{bmatrix}$

and thus can be written in a new form as

$$(\underline{\beta}_1^1, \underline{\beta}_2^1)\ \underline{q} = 0$$

where

$$\underline{q}^1 = (1, \gamma, -1, -\gamma).$$

In order to present an equivalent expression for the unconstrained least squares estimates, we define a new variable

$$\underline{F} = \hat{\underline{\beta}}_{LLs}\underline{q} \qquad \text{and} \quad \underline{C} = \underline{q}^1\ \underline{C}^{-1}\underline{q}$$

where

$$\underline{C}^{-1} = \begin{bmatrix} \underline{C}_1^{-1} & \underline{0} \\ \underline{0} & \underline{C}_2^{-1} \end{bmatrix}$$

Using equation (10) in the Appendix the constrained least squares estimate is given by

$$\hat{\underline{\beta}}_c = \hat{\underline{\beta}}_{LLs} - \underline{C}^{-1}\underline{q}F^2/G \tag{3.2.1}$$

The corresponding residual sum of squares is given by the sum of the individual residual sums of squares, from the unconstrained least squares estimation, and from the extra sum of squares resulting from incorporating the constraint. This is given by

$$RSS_i(\gamma) = \|\underline{Y}_1 - \underline{X}_1\hat{\underline{\beta}}_1(i)\|^2 + \|\underline{Y}_2 - \underline{X}_2\underline{\beta}_2(i)\|^2 + F^2/_G \tag{3.2.2}$$

(compare with equation (11) in the Appendix) where the first two terms are the individual residual sums of squares for the local least squares estimation. Substituting for the elements of F and G in (3.2.2).

$$F^2/_G = \frac{[\hat{\beta}_{11}(i) + \hat{\beta}_{12}(i)\gamma - \hat{\beta}_{21}(i) - \hat{\beta}_{22}(i)\gamma]^2}{\frac{1}{i} + \frac{1}{n-1} + \frac{(\gamma-\bar{x}_1(i))^2}{S_{x_1x_1}(i)} + \frac{(\gamma-\bar{x}_2(i))^2}{S_{x_2x_2}(i)}}$$

where $S_{x_jx_j}(i)$ $j = 1,2$ is the corrected sum of squares of the x-values for the data subset j for partition i, and $\bar{x}_j(i)$ denotes the corresponding mean x-value for the j^{th} subset and partition i. Constrained least

squares estimation is particularly useful when the join-points in segmented regression models are known. It is equally useful in the algorithm to be developed in the next Section.

In Chapter 2 we talked about $RSS(\gamma)$ which is the residual sum of squares function expressed as a function of the join-points. We have also defined another residual sum of squares obtained from constrained estimation denoted $RSS_i(\gamma)$. The two are related by

$$RSS(\gamma) = RSS_i(\gamma), \text{ if } \gamma \varepsilon [x_i, x_{i+1}]$$

Otherwise

$$= RSS_{i(\gamma)}(\gamma) \text{ where } i(\gamma) = \max \{i : x_i < \gamma\}$$

The calculation of $RSS(\gamma)$ or $RSS_{i(\gamma)}$ depends on whether local least squares estimation or constrained estimation is required. The form required followed accordingly.

3.2.1 General discussion of the algorithms

Before reviewing the algorithms used for solving the two segment model problem mentioned briefly in Section 3.1, some features possessed by these algorithms are discussed below.

The algorithms make exhaustive examination of all possible partitions of the given set of data into two subsets to which the segments are fitted. Efficient implementation thus involves the use of updating facilities in which the need to recalculate completely the necessary sums of squares, or cross products, for the subsets of each partition, is avoided. For instance, suppose we consider a partition of the observations into two subsets denoted $(x_1, y_1) \ldots . (x_i, y_i)$ and $(x_{i+1}, y_{i+1}) \ldots .. (x_n, y_n)$, using the updating facilities in fitting two segments model to the data. Consider partition i, the facility involves adding (or subtracting) the appropriate x-values at the point (that is x_i, x_i^2, $x_i y_i$) to (from) the summation achieved in the calculation for partition (i-1)

to values of the first subset (second subset). The advantages of the facility are:-

(i) Savings on computer time or labour if data are analysed manually.

(ii) Simple to use, adopt and implement.

The algorithms have the features that their implementation is easy to understand. In order to make any algorithms appealing to intending users, it is essential that such an algorithm possess the above qualities. This enables users not familiar with, or who have no access to, computer facilities, to use the algorithm with minimum difficulty when the need arises.

3.3 Theorem for 2-Segment (Linear) Segmented Regression Problem

Hudson studies the problem of a K-segmented regression model with polynomial segments as discussed earlier. He, however, made a classification of the types of join-points that could occur in any segmented regression problem. The types have been broadly classified into two groups, for the purposes of this Dissertation.

Firstly, the estimated join-point $\hat{\gamma}$ may lie strictly between two adjacent observations; that is, for some i, $x_i < \hat{\gamma} < x_{i+1}$; or, secondly, the estimated join-point may coincide with one of the observations - for some i, $\hat{\gamma} = x_i$. Generally, the statistician has no knowledge of which type of join-point is present in any data set. Hence the necessity to investigate all the possibilities in any analysis. We shall now consider the variants of Hudson theorems (1 and 2) which are restricted to the 2-linear segment case.

Theorem

(i) If the estimated point of intersection of the local least squares fitted lines, $\hat{\gamma}_{LLs}(i)$ lies in the interval $[x_i, x_{i+1}]$, $\hat{\gamma}_{LLs}(i)$ is the minimizing value (for the

residual sum of squares) of the join-point in that interval. Otherwise,

(ii) the minimizing value is x_i or x_{i+1}. Mathematically expressed:

(a) If $\hat{\gamma}_{LLs}(i) \in [x_i, x_{i+1}]$ then

$\underset{x_i<\gamma<x_{i+1}}{\text{Min}}$ $RSS_i(\gamma)$ is given by $\hat{\gamma} = \hat{\gamma}_{LLs}(i)$

(b) If $\hat{\gamma}_{LLs}(i) \notin [x_i, x_{i+1}]$ then

$\underset{x_i<\gamma<x_{i+1}}{\text{Min}}$ $RSS_i(\gamma)$ is achieved either by $\hat{\gamma} = x_i$ or by $\hat{\gamma} = x_{i+1}$

Proof

To prove the above theorem, it is required to be shown that the function $RSS_i(\gamma)$ has no more than one local minimum. Using the result given in equation (3.2.2)

$$RSS_i(\gamma) = \|\underline{Y}_1 - \underline{X}_1\hat{\underline{\beta}}_1(i)\|^2 + \|\underline{Y}_2 - \underline{X}_2\hat{\underline{\beta}}_2(i)\|^2 + \frac{[\hat{\beta}_{11}(i) + \hat{\beta}_{12}(i)\gamma - \hat{\beta}_{21}(i) - \hat{\beta}_{22}(i)\gamma]^2}{\frac{1}{i} + \frac{1}{n-i} + \frac{[\gamma-\bar{x}_1(i)]^2}{S_{x_1x_1}} + \frac{[\gamma-\bar{x}_2(i)]^2}{S_{x_2x_2}}}$$

The first two terms do not involve γ representing the local least squares residual sums of squares. The above can be simplified further to become

$$RSS_i(\gamma) = g + [a\gamma^2 + b\gamma + c]/[d\gamma^2 + e\gamma + f] \qquad (3.3.1)$$

where

$$g = \|\underline{Y}_1 - \underline{X}_1\hat{\underline{\beta}}_1(i)\|^2 + \|\underline{Y}_2 - \underline{X}_2\hat{\underline{\beta}}_2(i)\|^2$$

$$a = i(n-i)\, S_{x_1x_1}\, S_{x_2x_2}\, [\hat{\beta}_{12} - \hat{\beta}_{21}]^2$$

$$b = 2i(n-i) S_{x_1x_1}\, S_{x_2x_2}\, [\hat{\beta}_{11} + \hat{\beta}_{21}][\hat{\beta}_{12} - \hat{\beta}_{22}]$$

$$c = i(n-i) S_{x_1x_1}\, S_{x_2x_2}\, [\hat{\beta}_{11} - \hat{\beta}_{21}]^2$$

$$d = i(n-i)[S_{x_1x_1} + S_{x_2x_2}]$$

$$e = -2i(n-i)\ [\bar{x}_1 S_{x_2x_2} + \bar{x}_2 S_{x_1x_1}]$$

$$f = i(n-i)\ [\bar{x}_1^2 S_{x_1x_1} + \bar{x}_2^2 S_{x_2x_2}] + nS_{x_1x_1}\ S_{x_2x_2}$$

(Note that the subscript (i) has been dropped in the new form to avoid very clumsy functions only. The function is valid for all partitions i.) Differentiating $RSS_i(\gamma)$ with respect to γ we obtain

$$\frac{\partial RSS_i(\gamma)}{\partial \gamma} = \frac{[2a\gamma + b][d\gamma^2 + e\gamma + f] - [2d\gamma + e][a\gamma^2 + b\gamma + c]}{[d\gamma^2 + e\gamma + f]^2}$$

and setting the derivatives equal to zero gives

$$[ae - bd]\hat{\gamma}^2 + 2[af - cd]\hat{\gamma} + [bf - ec] = 0 \qquad (3.3.2)$$

The above equation is quadratic in $\hat{\gamma}$ and has two real roots one of which ($\hat{\gamma}_{LLs}(i)$) corresponds to a minimum, the other corresponding to a maximum. Since $RSS_i(\gamma)$ has at most one minimum, the point $\hat{\gamma}_{LLs}(i)$ provides the minimum residual sum of squares in the range $[x_i, x_{i+1}]$ if it lies within the range. Otherwise the minimum is provided at one of the endpoints, that is x_i or x_{i+1}. The theorem above forms the basis on which the improved algorithm, proposed in Section 3.4, is developed. This underlying theorem is also true for Hudson's and Hinkley's algorithms.

3.3.1 Description of Hudson's Algorithm

Given n data points whose abscissae are ordered, to which a continuous model with two-linear segments is to be fitted with no knowledge of the location of the join-point; the algorithm considers all possible partitions of the data into two disjoint subsets, corresponding to segments. It proceeds first to search for partition i, $(2 \le i \le n-2)$, in which the point of intersection of the local least squares fitted lines lies in the interval $[x_i, x_{i+1}]$. For such partitions, the local least squares estimates and the corresponding residual sums of squares are recorded. Those estimates are regarded as possible candidates for the final solution of the problem.

The algorithm proceeds to consider, as a second stage, all those partitions where the point of intersection $\hat{\gamma}_{LLs}(i)$ of the local least squares fitted lines does not lie in the interval $[x_i, x_{i+1}]$. Constrained fits are performed for those partitions such that the lines intersect at either x_i or x_{i+1}. Then the corresponding constrained estimates of the parameters and the residual sums of squares are also recorded for such partitions. The results are all tabulated. Finally, the algorithm chooses the final solution which provides the smallest residual sum of squares from the tabulated results for both constrained and unconstrained estimation.

3.3.2 Description of Hinkley's Algorithm

Hinkley's approach performs the two stages of Hudson's approach simultaneously. His algorithm considers all partitions i. For each i, $\gamma_{LLs}(i)$ is obtained. If it lies in the interval $[x_i, x_{i+1}]$, the local least squares fit solution and the residual sum of squares are recorded. Otherwise, constrained fit is performed for the two segments to intersect at x_i and x_{i+1}, with the corresponding parameter estimates and residual sums of squares recorded. The results are similarly tabulated. The final result is given by the solution providing the smallest residual sum of squares.

The algorithm when compared to Hudson's has some advantages. Firstly, it avoids performing the estimation as a two-stage process, hence saving computer time. Secondly, the updating facilities mentioned in Section 3.2.1 are easily effected in Hinkley's. The facilities are in operation in Hudson's though the process is repeated for each of the stages, at least for the partitions where the point of intersection of the local least squares fitted lines does not belong to the right interval. Hinkley's can be regarded as an improvement on Hudson's because of the

time-savings achieved.

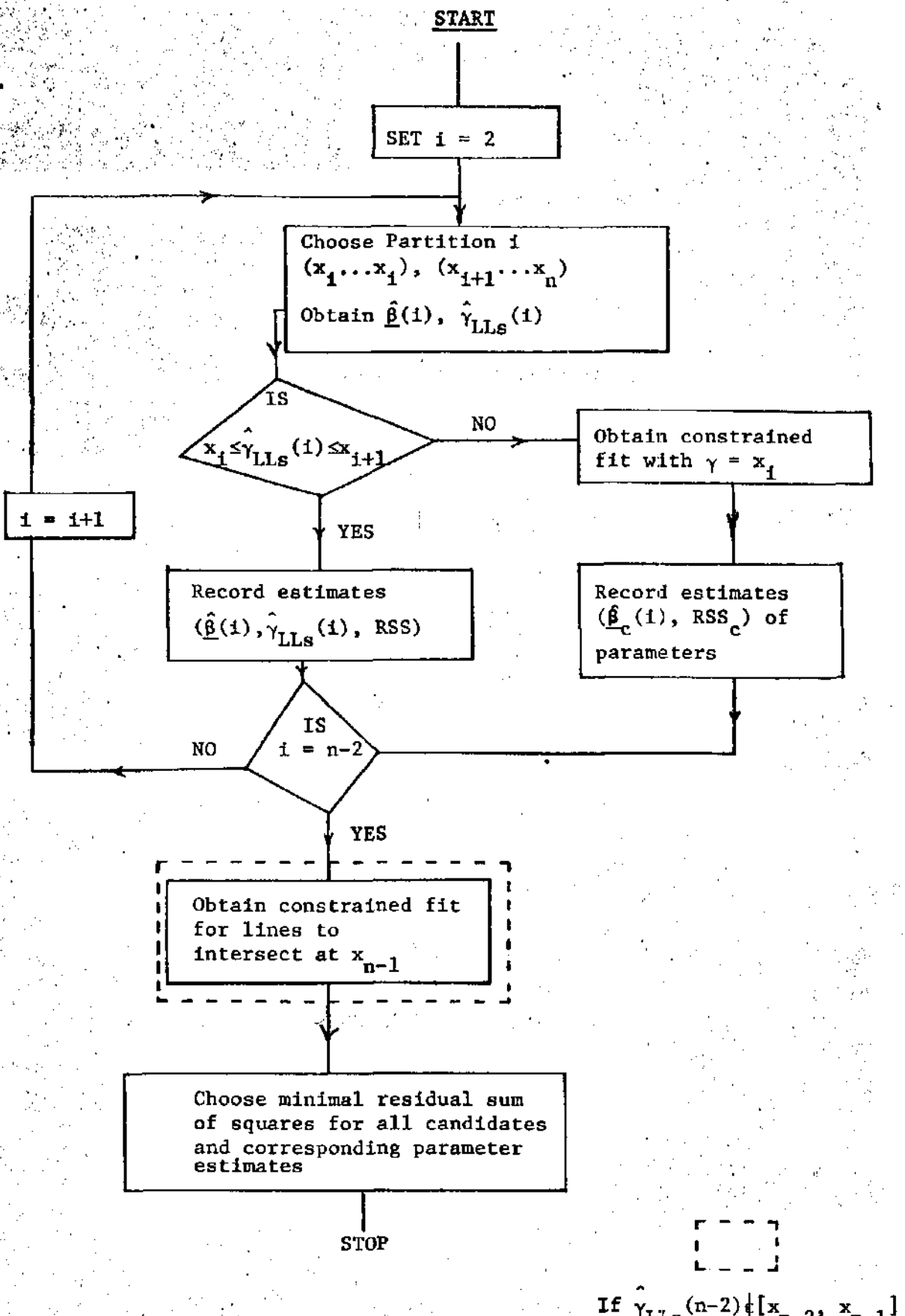
As a general comment, it seems desirable to carry out the search simultaneously, as did Hinkley. However, his method involves duplicated construction of constrained fits at x_i, x_{i+1} whenever the fitted local least squares lines intersect at wrong intervals for two consecutive partitions. Since the partition is in sequence, with i increasing by one point at a time, it is obvious that the possibility of performing a constrained fit with $\hat{\gamma} = x_{i+1}$ remains when the next partition is considered, and when it is found necessary. Apart from computer time, space may be a problem for large data sets when constrained fits are performed at both endpoints of the intervals x_i and x_{i+1} whenever $\hat{\gamma}_{LLs}(i)$ lies outside the interval. For a user applying the algorithm manually, it may be very laborious in terms of computations.

3.4 The Improved Algorithm

The improved algorithm is proposed for reasons discussed at the end of the previous section. The algorithm carries out the two stages simultaneously and also avoids the duplication of the constrained fits at points x_i and x_{i+1}, for any consecutive partitions for which $\hat{\gamma}_{LLs}(i)$ lies outside the interval $[x_i, x_{i+1}]$.

The algorithm is based like the others on the theorem given in Section 3.3. A flowchart of the improved algorithm follows below in Fig. 3.4.1. The chart possesses some self-explanatory features which make detailed explanations unnecessary. Like the others, the algorithm considers all possible partitions of the data points into two disjoint subsets, with the restriction on i that there should be at least two data points in any segment. For each partition i, the local least squares solution is recorded whenever the point of intersection of the fitted local least squares lines lie in the correct interval, $[x_i, x_{i+1}]$. Otherwise, the

Figure 3.4.1 Flowchart for the Improved Algorithm



fitted lines are constrained to intersect at the point x_i. Hence the constrained fit is conducted at the left endpoint of the interval whenever $\hat{\gamma}_{LLs}(i)$ lies outside the interval $[x_i, x_{i+1}]$ for all partitions. For the interval $[x_{n-2}, x_{n-1}]$ constrained fits are conducted at both endpoints whenever the $\hat{\gamma}_{LLs}(n-2)$ lies outside the interval. The fits at these points are performed to cater for the possibility of the estimated join-point coinciding with either x_{n-2} or x_{n-1}, i.e. $\hat{\gamma} = x_{n-2}$ or x_{n-1}.

As with the other two algorithms, the results are tabulated for all possible candidates' solutions. The solution leading to the smallest residual sum of squares provides the final solution. The corresponding parameters estimates provide the solution for the segments. The improved algorithm also possesses the features discussed in Section 3.2.1.

3.5 Illustration of the Use of the Improved Algorithm

Using the flowchart is Section 3.4, a Basic program was written to show how the improved algorithm works, some examples of data were analysed to illustrate solutions involving the different types of join-point mentioned earlier. Some of the examples are constructed artificially whereas one uses real data. Unlike Hudson's, these examples did not involve equally spaced abscissae and very small sample sizes. For each example, the graph of RSS(γ) is presented to help illustrate the role of the fundamental theorem in Section 3.2 on which the algorithm is based.

Example 1

This example is constructed to illustrate the situation in which the estimated join-point lies strictly between two adjacent data points.

Table 1

X	2.4	3.4	4.4	5.4	6.4	7.0	8.6	9.8	11.6	12.6	13.8
Y	11.6	8.8	7.2	5.2	3.2	2.0	5.2	6.8	9.6	12.0	14.0

The graph for the above data is shown in Fig. 3.5.1a below. It is obvious

GRAPH OF DATA POINTS

X	2.4	3.4	4.4	5.4	6.4	7.0	8.6	9.8	11.6	12.6	13.8
Y	11.6	8.8	7.2	5.2	3.2	2.0	5.2	6.8	9.6	12.0	14.0

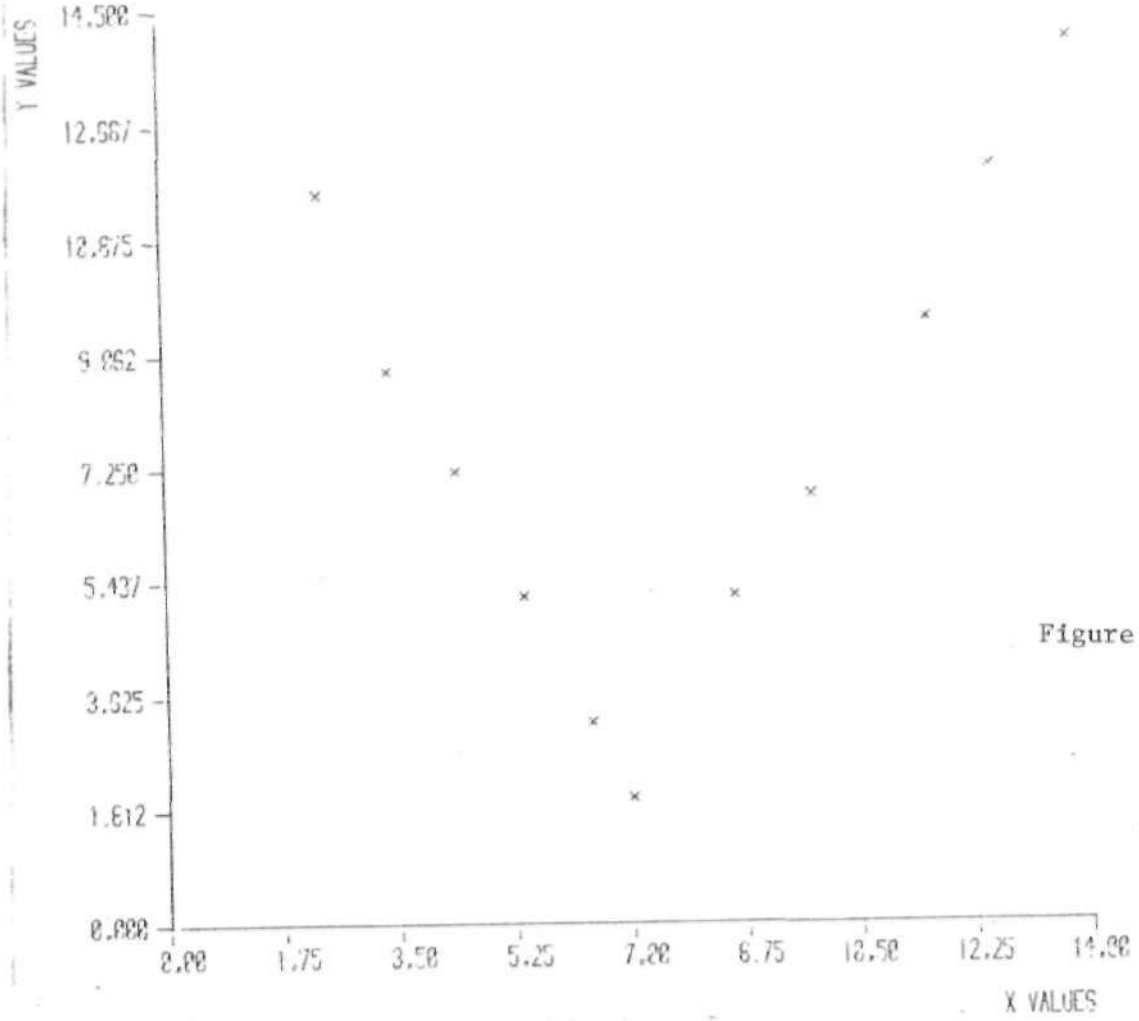


Figure 3.5.1a

GRAPH OF RSS(X) AGAINST JOIN POINT(X)

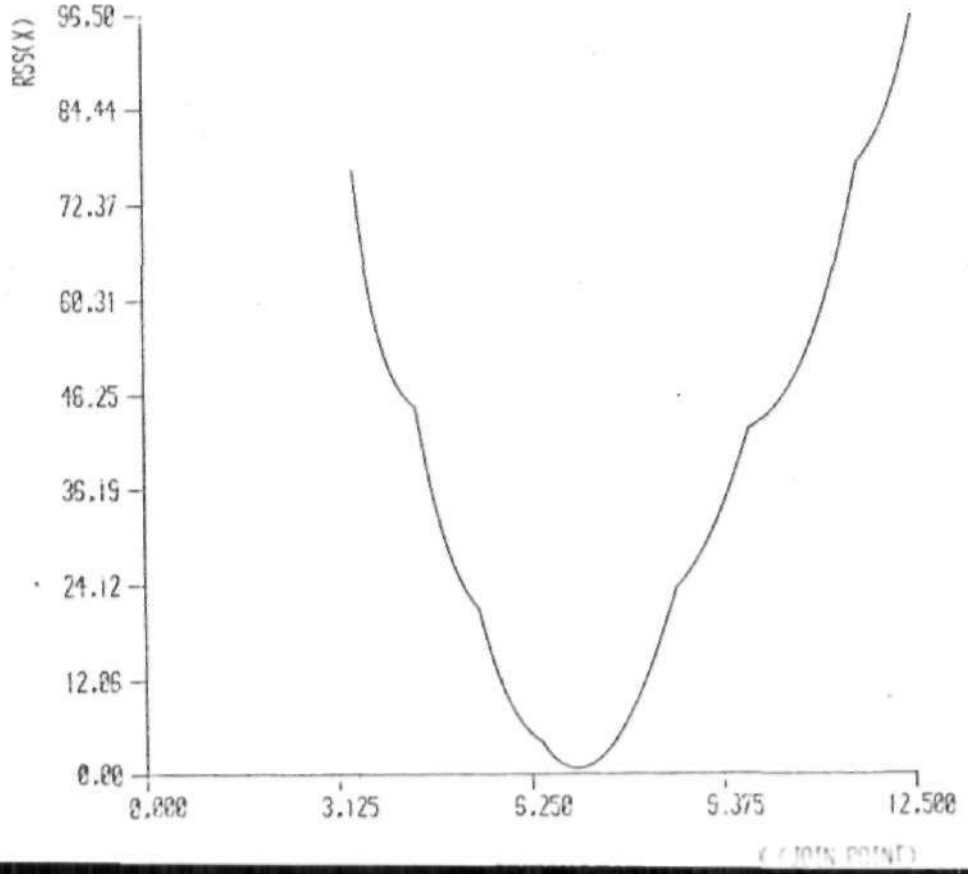


Figure 3.5.1b

from the graph that the data cannot be adequately represented by a single polynomial. The table below summarises the operation of the algorithm.

Table 1a Summary of Stages of the Algorithm

Partition (i)	Range for γ	$\gamma_{LLs}(i)$	γ	RSS(γ)	Remark type of estimation
2	$3.4 \leq \gamma < 4.4$	5.204	3.4	76.6913	Constrained
3	$4.4 \leq \gamma < 5.4$	6.167	4.4	46.612	Constrained
4	$5.4 \leq \gamma < 6.4$	6.695	5.4	20.965	Constrained
5	$6.4 \leq \gamma < 7.0$	6.953	6.53	0.724	Local least squares
6	$7.0 \leq \gamma < 8.6$	6.941	7.0	0.739	Constrained
7	$8.6 \leq \gamma < 9.8$	7.826	8.6	23.628	Constrained
8	$9.8 \leq \gamma < 11.6$	8.789	9.8	43.753	Constrained
9	$11.6 \leq \gamma \leq 12.6$	9.048	(11.6	77.522	Constrained
			(		
			(12.6	1100.251	Constrained

We shall consider the table; partition i is the grouping of the data set into two disjoint subsets, subject to the restriction that each contains at least two data points. The term "range" for γ denotes the region in which possible values for the estimated join-point $\hat{\gamma}$ are being investigated, the range corresponding to partition i being $[x_i, x_{i+1}]$. γ column are candidates for $\hat{\gamma}$, and the RSS(γ) column tabulates the corresponding values for the residual sum of squares. In this algorithm constrained fit is always performed at the left endpoint of the range, with the exception of points x_{n-1} where constrained fit was performed at the two endpoints of the range x_{n-2}, x_{n-1} whenever it is necessary. The point is the same as the $\hat{\gamma}_{LLs}$ for a particular partition whenever the estimated $\hat{\gamma}_{LLs}$ lies in the range for the partition. Table 1a above and the graph of a plot of RSS(γ) against γ given in Fig. 3.5.1b will be

used along with the theorem in Section 3.3 to explain how the improved algorithm operates.

Consider the first partition of the data represented by 2 on the Table, the first subset contains the first two data points and the second subset contains the remaining points. The algorithm performs a local least squares fit and it is found that $\hat{\gamma}_{LLs}(2)$ does not lie in the interval (3.4, 4.4). $\hat{\gamma}_{LLs}(2)$ lies to the right of the point 4.4, ($\gamma_{LLs}(2) = 5.204$). By the underlying theorem given in Section 3.3, the minimizing value of the residual sum of squares is either at x_i or x_{i+1}. The value chosen is x_i by the improved algorithms argument given in Section 3.4. As can be seen from the table, a constrained fit was performed at the left endpoint (i.e. $x_i = 3.4$) of the range.

Let us look at the behaviour of RSS(γ) within this interval (3.4, 4.4). It can be noticed from Fig. 3.5.1b that RSS attained its minimum within the range at (4.4). However, the algorithm has chosen to perform constrained fit at 3.4 instead of 4.4. Logically, given that $\hat{\gamma}_{LLs}(2)$ lies to the right of the right endpoint of the range under consideration, a constrained fit might be expected at a point to the direction of the minimum. However, as we shall see later, an algorithm based on such conjecture does not always prove reliable. Using the argument presented for proposing the improved algorithm, the constrained fit is to be always performed at the left endpoint of the range whenever $\hat{\gamma}_{LLs}(i)$ does not lie within that range.

The next partition is then considered and similarly, it is found that $\hat{\gamma}_{LLs}(3)$ lies to the right of the interval (4.4, 5.4) and thus a constrained fit is performed for the lines to intersect at 4.4. The behaviour of the RSS(γ) should be noticed accordingly along with the underlying theorem. The process is continued as we consider the other partitions. Note that for partition 5, in the Table, $\hat{\gamma}_{LLs}(5)$ lies in the

correct range and the local least suqares solutions are recorded accordingly by the theorem. As we proceed for other partitions beyond this point, the value of RSS(γ) starts to rise (Graph 3.5.1b). One important point as we consider partitions 6, 7, 8 is that the points of intersection of the fitted local least squares lines are observed to lie to the left of the range which may be considered as an indication that the region where the minimum lies has been exceeded. As already noted from the example given in Section 2.2, the gradient discontinuities of RSS(γ) at the observed x-values are evident. Also, the function RSS(γ) has continuous derivatives between any two observations.

As indicated on the flow chart for the improved algorithm, constrained fit is performed at both endpoints of the range for the last partition (11.6, 12.6) since $\gamma_{LLs}(9)$ does not lie in the range. The results are as given on the Table except the parameter estimates are not all included there. The smallest attainable residual sum of squares is obtained by a local least squares fit for partition 5 with parameter estimates given by

$$\hat{\underline{\beta}}_1 = (16.176, \ -2.040)$$

$$\hat{\underline{\beta}}_2 = (-10.082, \ 1.736)$$

$$\hat{\gamma} = 6.953.$$

The residual sum of squares is 0.724.

Example 2

This example illustrates the case when the estimated join-point coincides with one of the data points.

Table 2

X	1.0	2.0	3.5	4.0	4.5	6.0	9.0	10.0	10.5	11.0
Y	10.0	9.6	6.5	6.2	5.8	2.8	11.0	12.0	13.2	13.6

The graph of the data is given in Fig. 3.5.2a. As for the previous

example, the table below summarises the procedures of the algorithms while Fig. 3.5.2b gives the behaviour of RSS(γ).

Table 2a

Partition i	Range for γ	$\hat{\gamma}_{LLs}(i)$	γ	RSS(γ)	Remark type of estimation
2	$2.0 \leq \gamma < 3.5$	6.3875	2.0	56.1207	Constrained
3	$3.5 \leq \gamma < 4.0$	4.6457	3.5	29.0622	Constrained
4	$4.0 \leq \gamma < 4.5$	5.1206	4.0	23.7477	Constrained
5	$4.5 \leq \gamma < 6.0$	6.09332	4.5	17.7736	Constrained
6	$6.0 \leq \gamma < 9.0$	4.6937	6.0	2.74706	Constrained
7	$9.0 \leq \gamma < 10$	6.9032	9.0	52.5756	Constrained
8	$10.0 \leq \gamma \leq 10.5$	3.4623	(10.0	65.9046	Constrained
			(10.5	72.6245	Constrained

The above table is self-explanatory but we shall similarly relate it and the graph (Fig. 3.5.2b) of RSS(γ) over the entire range using the underlying theorem on which the algorithm is based. The arguments and explanation presented in the previous example hold here also. The major difference in this example is that, the minimum obtained is attained by constrained least squares fit.

A look at the table above showed that none of the local least squares fitted lines intersect at the correct range for all partitions. So constrained fits are performed over the entire range. The final solution is chosen from the set of candidates. The minimum was attained when the lines were constrained to intersect at $x_i = 6$. At this point, notice the occurrence of gradient discontinuity. Note also the direction of the minimum value of the residual sum of squares as we proceed from partitions 2 to 5. In partitions i, such that $2 \leq i \leq 5$, $\hat{\gamma}_{LLs}(i)$ is always to the right

GRAPH OF DATA POINTS

X	1.0	2.0	3.5	4.0	4.5	6.0	9.0	10.0	10.5	11.0
Y	10.0	9.6	6.5	6.2	5.8	2.8	11.0	12.0	13.2	13.6

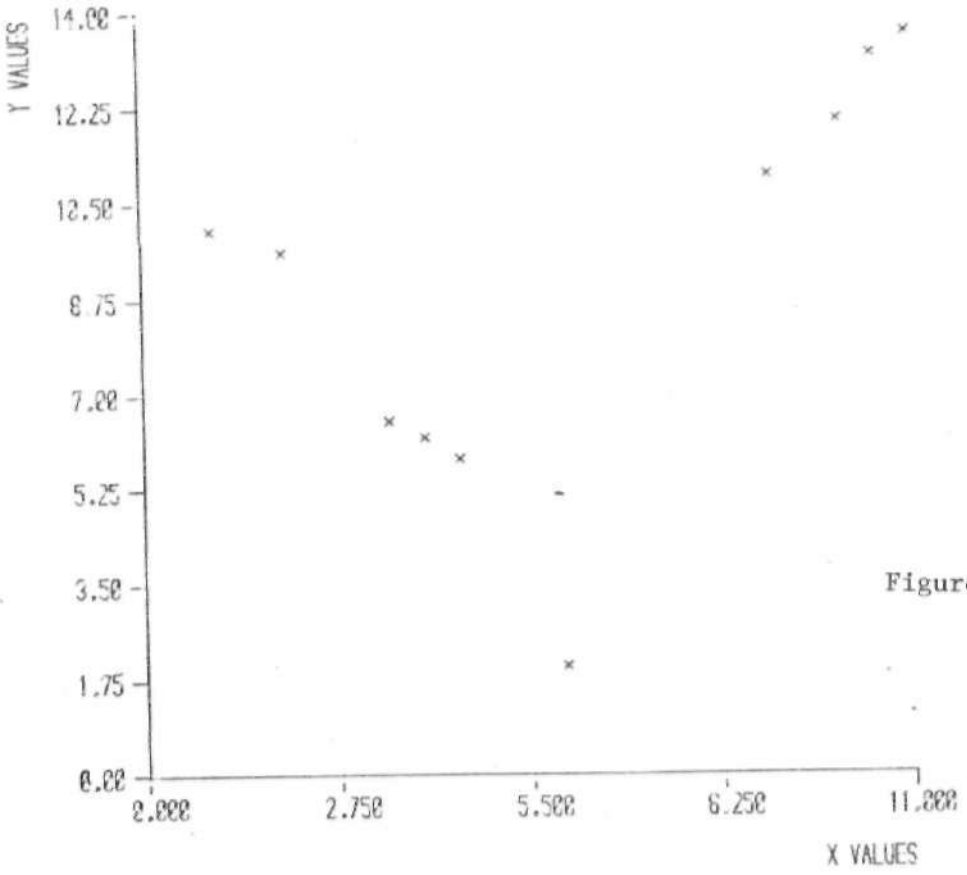


Figure 3.5.2a

GRAPH OF RSS(X) AGAINST JOIN POINT(X)

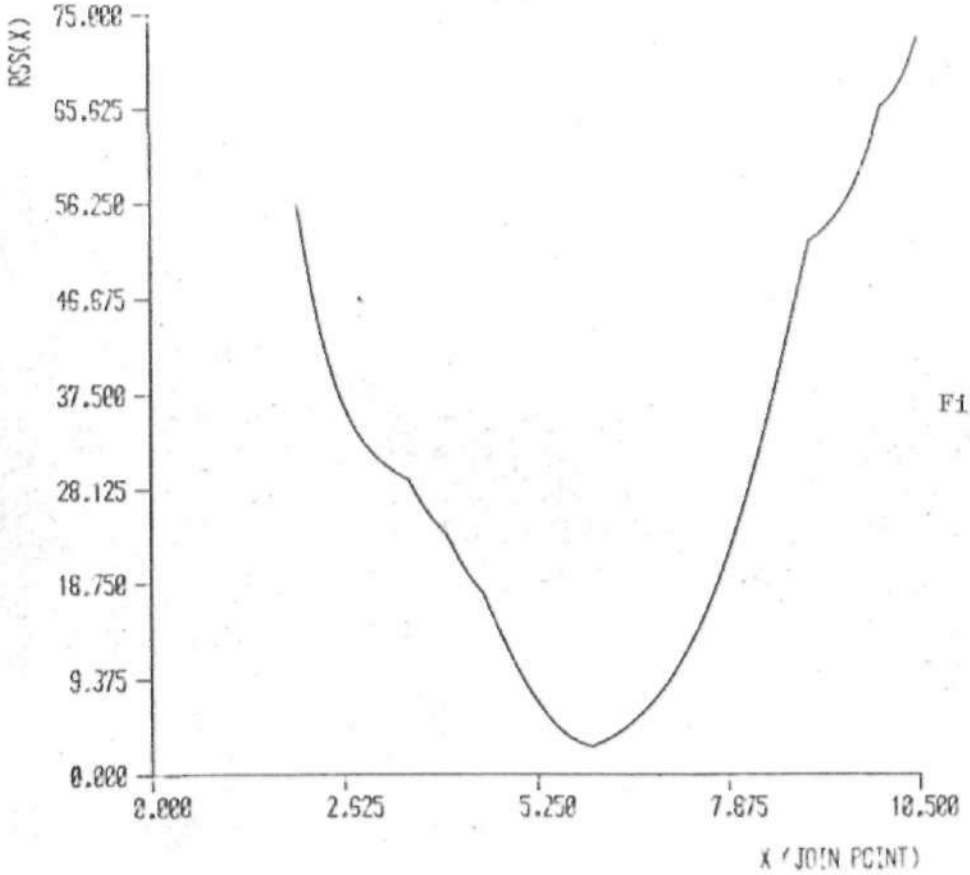


Figure 3.5.2b

of the interval (x_i, x_{i+1}). But for partitions i, such that $6 \leq i \leq 8$, the location predicted by $\hat{\gamma}_{LLs}(i)$ switches over to the left of the range, implying that the region, where the minimum lies, is exceeded, as notice in Example 2. As suggested earlier, these changes could be useful but should definitely not be used as a basis for an algorithm.

The solution of the parameters providing the smallest residual sum of squares is given by

$$\hat{\underline{\beta}}_1 = (20, \ -2)$$

$$\hat{\underline{\beta}}_2 = (-12, \ 2)$$

$$\hat{\gamma} = 6.0$$

and RSS = 2.7471.

Example 3

This Example involves real data taken from the paper by Beckman and Cook (1979). The data were obtained from an industrial experiment on yield of hydrogen sulphide (response) from gamma radiolysis as a function of the Krypton pressure (independent variable).

Table 3

X	1.00	1.50	1.75	2.10	2.35	2.65	3.00
Y	6.40	7.00	7.40	8.80	9.00	6.40	6.60

The plot of the data above is similarly given in Fig. 3.5.3a. Looking at the graph, it is not immediately clear whether the use of a polynomial will adequately represent the relationship between Y and X instead of the segmented linear regression. The first five points appear to suggest some linear relationship, while the last two points clearly stand out from the rest. Segmented regression techniques were applied to model the relationship. To avoid much repetition, we shall only display

Example 1 (Data from Beckman and Cook (1979))

X	1.00	1.50	1.75	2.10	2.35	2.65	3.0
Y	6.4	7.0	7.4	8.8	9.0	6.4	6.6

GRAPH OF DATA POINTS

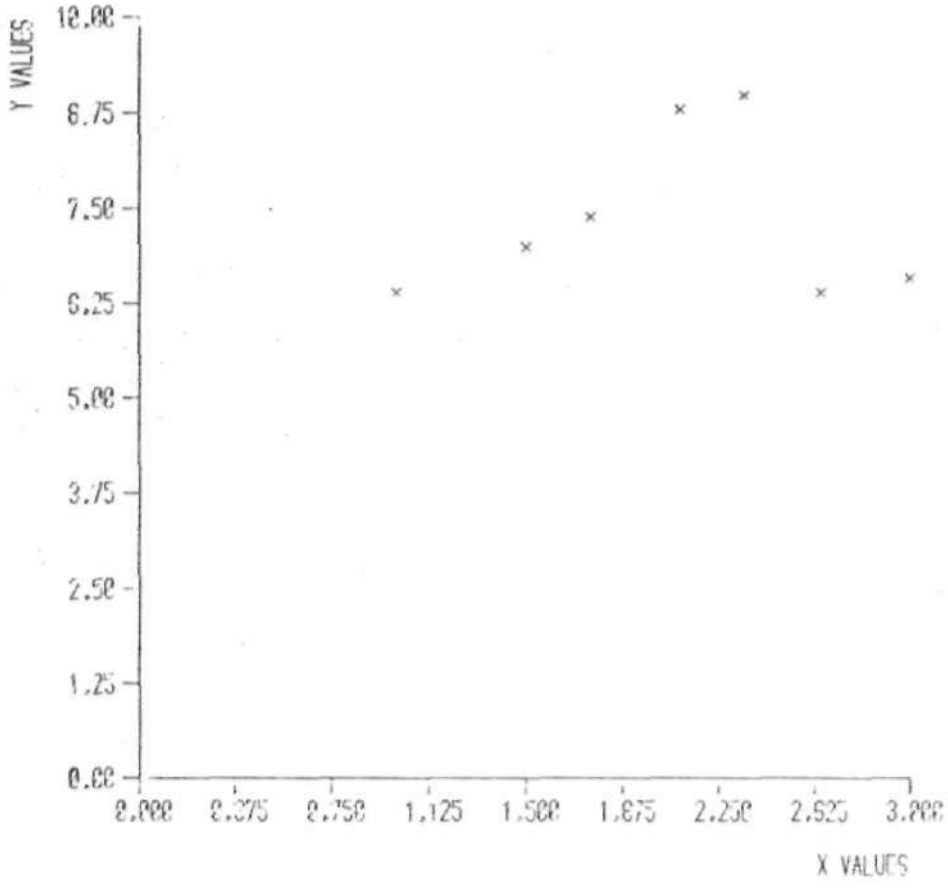


Figure 3.5.3a

GRAPH OF RSS(X) AGAINST X

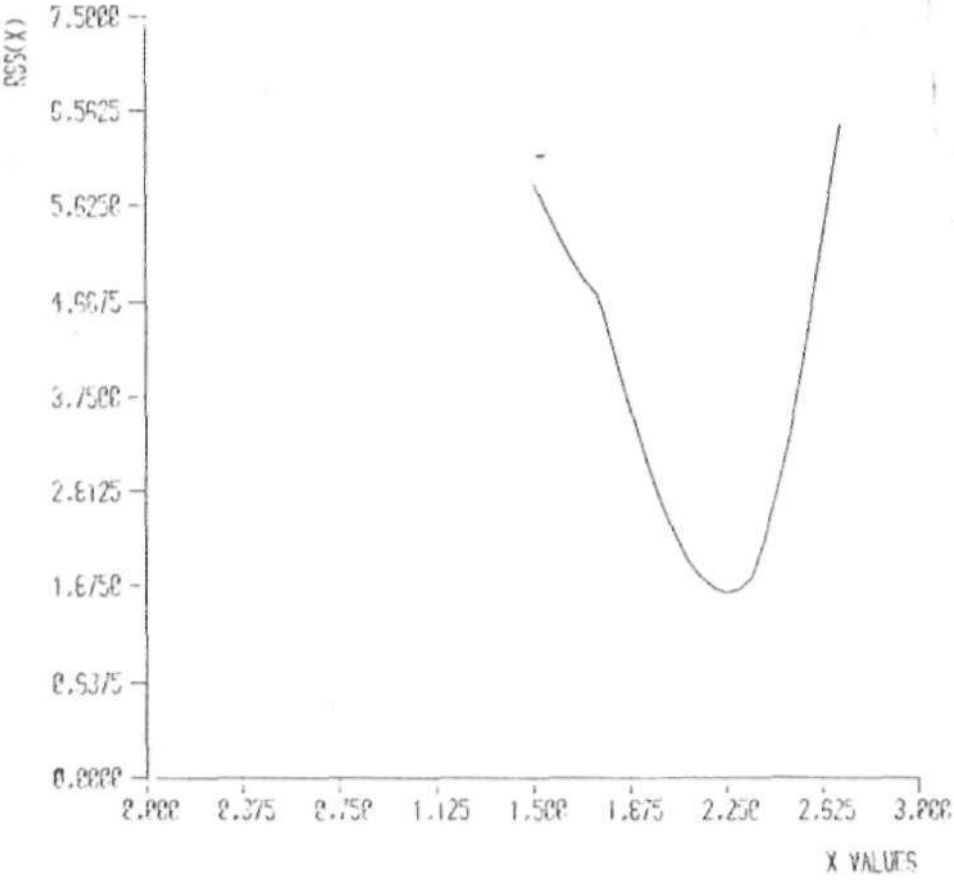


Figure 3.5.3b

the summary table, the parameter estimates as well as the graph showing the form of RSS(γ) for set of trial values of γ.

Table 3a Summary

Partition i	Range for γ	$\gamma_{LLs}(i)$	γ	RSS(γ)	Remark type of estimation
2	$1.5 \leq \gamma < 1.75$	2.2071	1.5	5.873	Constrained
3	$1.75 \leq \gamma < 2.10$	2.3669	1.75	4.757	Constrained
4	$2.10 \leq \gamma < 2.35$	2.2570	2.2570	1.804	Local least squares
5	$2.35 \leq \gamma \leq 2.65$	0.5300	(2.35	1.945	Constrained
			(2.65	5.619	Constrained

The graph of RSS(γ) for all values of i is given in Fig. 3.5.1b. The final solution is given with the first segment fitted to the subset containing the first four observations, and the second segment fitted to the subset containing the remaining points. The final solution is as follows:

$$\hat{\underline{\beta}}_1 = (4.085, 2.088)$$

$$\hat{\underline{\beta}}_2 = (16.887, -3.575)$$

$$\hat{\gamma} = 2.257$$

giving the rise to RSS = 1.804. For these data, the solution is obtained by local least squares estimation. The above result was found to be the same as that obtained by the authors.

3.6 Further Improvements and General Discussion on Algorithm

The improved algorithm could be improved further by incorporating a slight refinement. The refinement will be to avoid performing constrained fits with $\gamma = x_i$, when dealing with the interval (x_i, x_{i+1}), if it was found that the point of intersection of the fitted local least lines of the previous partition lies in the correct interval. This is so

because by the theorem underlying the algorithm, it shows that there is no point at which a lower residual sum of squares is obtainable when $\hat{\gamma}_{LLs}(i)$ lies in the correct range. Another improvement that could be incorporated will be the point raised by Hudson. He remarked that a constrained fit is no longer necessary for partition i whenever it is found that the local least squares residual sum of squares for the partition is greater than the smallest residual sum of squares so far recorded in the table.

If the first refinement had been incorporated in our algorithm, the constrained fit in partition 6 of Example 1 might not have been performed. But in that Example the results are close enough to make it very obvious. So although $\hat{\gamma}_{LLs}(6)$ lies outside the interval, the constrained fit will not have been necessary.

In Example 1, a brief mention was made of the danger of performing the constrained fit at the endpoint of the interval closer to the point of intersection of the local least squares lines (i.e. $\hat{\gamma}_{LLs}(i)$). In Fig. 3.5.1b, a look at the first range showed that RSS(4.4) is smaller than RSS(3.4). Essentially, within the interval (3.4, 4.4), the behaviour of RSS(γ) showed that the minimum is attained at the right endpoint. This seems to suggest that the constrained fit should have been most appropriate at $\gamma = 4.4$ instead of 3.4. The same argument holds for the next partition. It may be worthwhile to point out that algorithms which have the basis discussed above as their underlying principle would have worked for all the examples in this Dissertation, as well as all examples that have been published in literature. However, it must be stated, as shall be seen in Example 4 below, that such algorithms may produce incorrect results.

Example 4

This example is constructed to illustrate the danger of using an algorithm which is based on the conjecture that the constrained fit be

performed at the endpoint closer to the predicted point of intersection of the fitted local least squares lines.

Table 4

X	2.0	7.0	10.5	11.5
Y	2.0	4.0	2.5	3.0

The plot of the above data is given in Fig. 3.5.4a. The data constructed artificially consist of only four data points. The only permissible partition of the data is into subsets consisting of the points with abscissae (2.0, 7.0) and (10.5, 11.5). By local least squares estimation $\hat{\gamma}_{LLs}(2) = 39.5$. This point clearly is out of the range where interval is given by (7.0, 10.5). In this case, the conjecture would lead one to perform the constrained fit at 10.5 since it is closer to 39.5, the point of intersection of the fitted local least squares lines. This could mean the choice of $\gamma = 10.5$. However, this choice would obviously have been in error because a look at the form of $RSS(\gamma)$ in Fig. 3.5.4b suggests that the minimum residual sum of squares is attained at $\gamma = 7$ and not 10.5, which is clearly a contradiction.

Though the data may appear to be unrealistic, they are presented to emphasize the fact that the conjecture is not always reliable and should not be made the basis for an algorithm. The improved algorithm has no difficulty in obtaining the correct solution because it is not based on the conjecture, but on the theorem given in Section 3.3.

Finally, Hudson, however, has a result which is worthy of brief mention that indicates roughly that an algorithm, performing a constrained fit only in the direction of $\hat{\gamma}_{LLs}$, will work, provided $\hat{\gamma}_{LLs}$ lies reasonably close to the data.

Example 4

X	2.0	7.0	10.5	11.5
Y	2.0	4.0	2.5	3.0

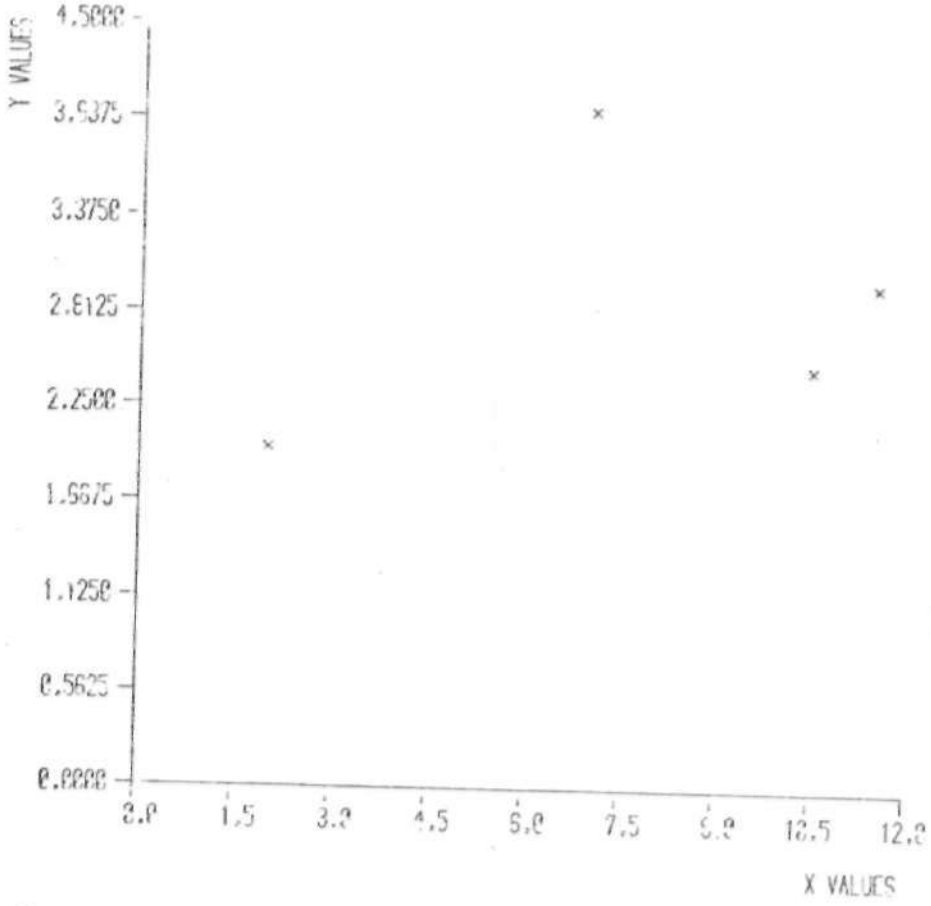


Figure 3.5.4a

GRAPH OF RSS(X) AGAINST JOIN POINT(X)

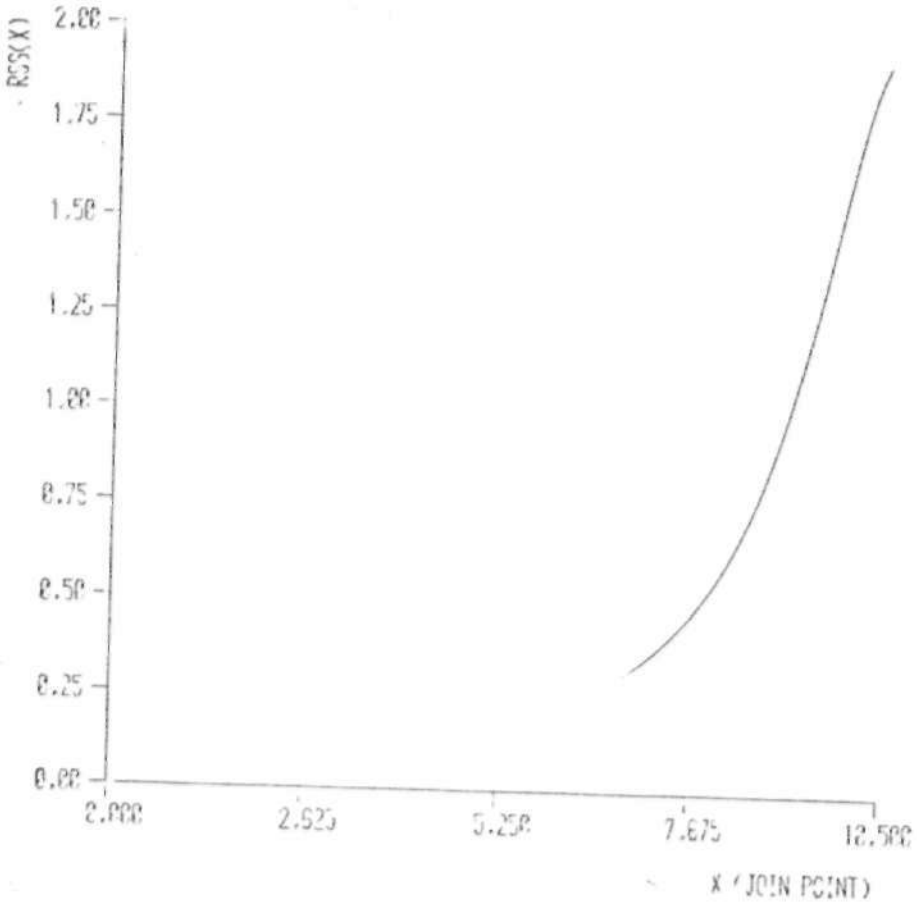


Figure 3.5.4b

4. OTHER ESTIMATION TECHNIQUES FOR CONTINUOUS MODELS

4.1 General Discussion

Up to this stage, the techniques, or algorithms, discussed are applicable to only two-segment problems. Specifically, consideration was given to linear segments. In this Chapter, we shall consider some procedures that are applicable to the more than two-linear segment model, defined in equation (2.2.1). Later, another technique which was used for solving the two-segment problem, and an underlying theory, different from the least squares approach, is considered.

Most of the basic definitions given for the two-segment problem can be extended, with slight modifications, to apply in the general case. Suppose there are K segments, there are (K-1) join-points denoted by the vector $\underline{\gamma}' = (\gamma_1, \gamma_2 \ldots . \gamma_{K-1})$. The corresponding residual sum of squares is given by $RSS(\underline{\gamma})$. This function plays an important role in one of the procedures to be considered in Section 4.3.

Hudson (1966) presented the general theory of estimations for the K-segment model where segments are polynomial functions. Williams (1970) suggested the extension of some of Hudson's ideas for solving the two-segment problem to deal with the three-segment situations. He derived the residual sum of squares function $RSS(\gamma_1, \gamma_2)$ which he later used to propose another algorithm of his own.

There is, however, a general difficulty associated with the estimation procedures, as already discussed in Section 2.2. Although this problem already exists, it becomes more pronounced where there is more than one join-point. Recall that $RSS(\underline{\gamma})$ has gradient discontinuity at the observed x-values for the two-segment problem. For the K=3 segments situation, $RSS(\underline{\gamma})$ is partially non-differentiable with respect to the vector of join-points $\underline{\gamma}$, whenever $\underline{\gamma}$ coincides with some observed x-values. Hudson

recommended some approximation techniques called "Successive Approximations". Applying the method to a three-segment problem, the procedure chooses a grid of values in the γ_1, γ_2 plane and then, using a method of steepest descent, an approximate solution is determined. Using the method, Hudson, for instance, remarked that RSS($\underline{\gamma}$), where $\underline{\gamma} = (\gamma_1, \gamma_2)$ is likely to possess several isolated differentiable local minima, where each local minimum corresponds to a case where γ_1 and γ_2 both lie between two adjacent data points.

We shall in the next section consider a preliminary analysis essential for a number of the estimation techniques, which suggests whether or not segmented regression procedures should be used for the available data. Then we proceed to look at a major technique called Grid Search, after which the Bayesian technique, which has a different underlying theory, will be considered. There are other methods which give approximate solutions for continuous segmented regression models. Among these is the "Search Algorithm" which was applied to segmented linear regression problems by Ertel and Fowlkes (1976). They imposed a restriction that the estimated join-points lie midway between two adjacent observed x-values. Bellmann and Roth (1969) used dynamic programming as a tool to approximate a polynomial by a segmented regression model.

4.2 Graphical Technique (Plotting)

The method involves the plotting of the response variable against the independent variable for the set of data. Data plotting is particularly useful in connection with segmented regression techniques for the following reasons.

Firstly, it is used in deciding whether or not a Segmented/Piecewise model is the most appropriate for the data. Recall that in Section 2.3 it was remarked that most of the applications of the technique were base on the graphical display of the data rather than any theoretical argument.

This implies that the decision to use either Segmented or Piecewise regression model was on the basis of the plots which provide a visual representation of the data required to be modelled.

Another reason for adopting the graphical procedure is that it enables the statistician to decide on the number of segments that appear to describe the data best. For a very large set of data, the graphical display of the data serves as a guideline to enable the partitioning of the data into disjoint sets to be performed around the region of the change from one segment to another. Furthermore, this provides a rough estimate of the region of change from one segment to another. This therefore assists in the formation of the initial partitioning.

The second point above is not essential in the use of the algorithms discussed in Chapter 3, because the algorithm is fairly automatic, provided it is already established that the segmented regression model is adequate. This implies that the first point remains relevant in connection with these algorithms. Most of the authors dealing with data analysis recommend that the procedure be applied for any analysis. It occasionally reveals some unanticipated features occurring in the data. Williams (1970) advocated that the data should always be plotted since it helps to reveal the regions of the change from one segment to another. Fischer et al.(1973) advocated them for all data analysis and in particular for Segmented/ Piecewise regression analysis. Other authors making a similar point are Gallant and Fuller (1973) and Beckman and Cook (1979).

4.3 Grid Search

The use of a Grid Search approach in solving segmented regression problems was demonstrated by Lerman (1980). The method is based on conducting a search, over a grid, of trial values of the vector of the join-points $\underline{\gamma}$. The corresponding residual sum of squares RSS($\underline{\gamma}$) is

calculated. The procedure then maps RSS($\underline{\gamma}$) against the vector $\underline{\gamma}$ by plotting the function against the vector. The form of RSS($\underline{\gamma}$) for the general K-segmented regression problem may be deduced from equation (11) in the Appendix. The form for the two-segment example derived in Section 3.2.2 was used in explaining the operation of the improved algorithm, along with some other results for the examples discussed earlier in Section 3.5. The grid values at which RSS($\underline{\gamma}$) is a minimum are chosen for the estimate of the vector of join-points. An iterative improvement, or refinement, may then be used to improve the final results (e.g. parameter estimates for the segments).

Grid Search has a major advantage over other estimation techniques in that it is widely applicable. It can be applied to two-segment or K-segment models, irrespective of whether each segment is linear or a polynomial. There are, however, some disadvantages associated with the technique. Firstly, when there are more than two join-points, the residual sum of squares function is no longer simple to present graphically, and the interpretation becomes more complex. To overcome this, Lerman recommended a procedure whereby the difficulty can be reduced. The second disadvantage is that it involves making an exhaustive search of all possible values of $\underline{\gamma}$ in order to obtain the form of RSS($\underline{\gamma}$). Hence, although it is economical in terms of space, it is not in terms of computer time. The problem of exhaustive search may be solved or reduced when the graphical procedure, discussed in Section 4.2, is adopted, because it may help to suggest a restricted region within which to conduct the Grid Search.

4.4 Bayesian Technique

This technique, applied to segmented regression problems was developed by Ba¢on and Watts (1971). They applied it to solve a two-linear segment continuous problem, whose model is given by:

$$E(Y) = \theta_0 + \theta_1(x-\gamma) + \theta_2(x-\gamma)\ Sig(x-\gamma) \qquad (4.4.1)$$

where θ_0 is the expectation of Y when $x = \gamma$, θ_1 denotes the average slope of the two segments, and where θ_2 is the difference between the slopes, and Sig(t) is a sign function defined as follows.

$$Sig(t) = \begin{cases} -1 & \text{if } t<0 \\ 0 & \text{if } t=0 \\ 1 & \text{if } t>0 \end{cases}$$

Model (4.4.1) is equivalent to the type discussed for the two-segment model in Chapter 2. The only difference is the fact that Model (4.4.1) has the continuity constraint imposed in the model, apart from the other differences for the parameter defined above.

The authors transformed Model (4.4.1) into a new one by replacing the Sig(t) function by a transition function given by $\tanh[t/\phi]$ where ϕ is the transition parameter; thus the model becomes

$$E(Y) = \theta_0 + \theta_1(x-\gamma) + \theta_2(x-\gamma)\ \tanh\ [(x-\gamma)/\phi] \qquad (4.4.2)$$

The transition function is chosen since it satisfied the following conditions:

(i) $\lim_{t \to \infty} \tanh[t/\phi] = 1$

(ii) $\lim_{t \to -\infty} \tanh[t/\phi] = -1$

(iii) $\lim_{\phi \to 0} \tanh[t/\phi] = Sig(t)$

The first and second conditions ensure that $\tanh[t/\phi]$ behaves like Sig(t) for large values of t. The third condition ensures that $\tanh[t/\phi]$ behaves like Sig(t) for small values of ϕ. Bacon and Watts remarked that there are other transition functions which satisfy the above conditions. Furthermore, they stressed that the form chosen is of secondary importance provided the conditions above are satisfied. The arguments are that, firstly, the effect of variability in the data will in

general mark any differences introduced by different transition functions. Secondly, that the general behaviour of the quantity $(x-\gamma)\tanh[(t-\gamma)/\phi]$ is only slightly affected by the particular form of $\tanh[t/\phi]$.

The introduction of the transition function leads the Model (4.4.2) to have an interesting feature making the technique worthy of consideration. Model (4.4.1) has discontinuous derivative at $x=\gamma$. The introduction of the transition function has transformed the model to one which possesses continuous derivative at this point. This change is best appreciated if we consider the stages of transformation by looking at the plot of the transition function, $t\tanh[t/\phi]$ against t taken from the authors' paper. The graph below illustrates how the transformation has come about for various choices of the transition parameter. The authors gave no reason for restricting the transition parameter to take values within the range $0\leq\phi\leq1$. It is presumed that ϕ could take values possibly greater than $1(\phi>1)$ leading to an even smoother change.

Taking a closer look at the graph, it can be observed that there is a gradient discontinuity at $\phi=0$ and this is gradually transformed such that the model possesses some continuous derivative for all $\phi=1$. Hence, using the model, the parameters are estimated by the Bayesian approach, where the parameters are θ_0, θ_1, θ_2, the join-point γ, the transition parameter ϕ, the variance σ^2.

Bacon and Watts remarked that the idea of transition functions could be incorporated in order to solve the segmented regression problems with $K\geq2$. Presumably, this may be achieved if some appropriate transition functions are introduced, because it was not stated specifically how this can be achieved. This could lead to a highly complicated model, which is one difficulty in using the technique.

Finally, a general comment about the technique is that, apart from the pioneers, no other users have been found in the literature. Furthermore,

Plot of $t \tanh(t/\phi)$ against

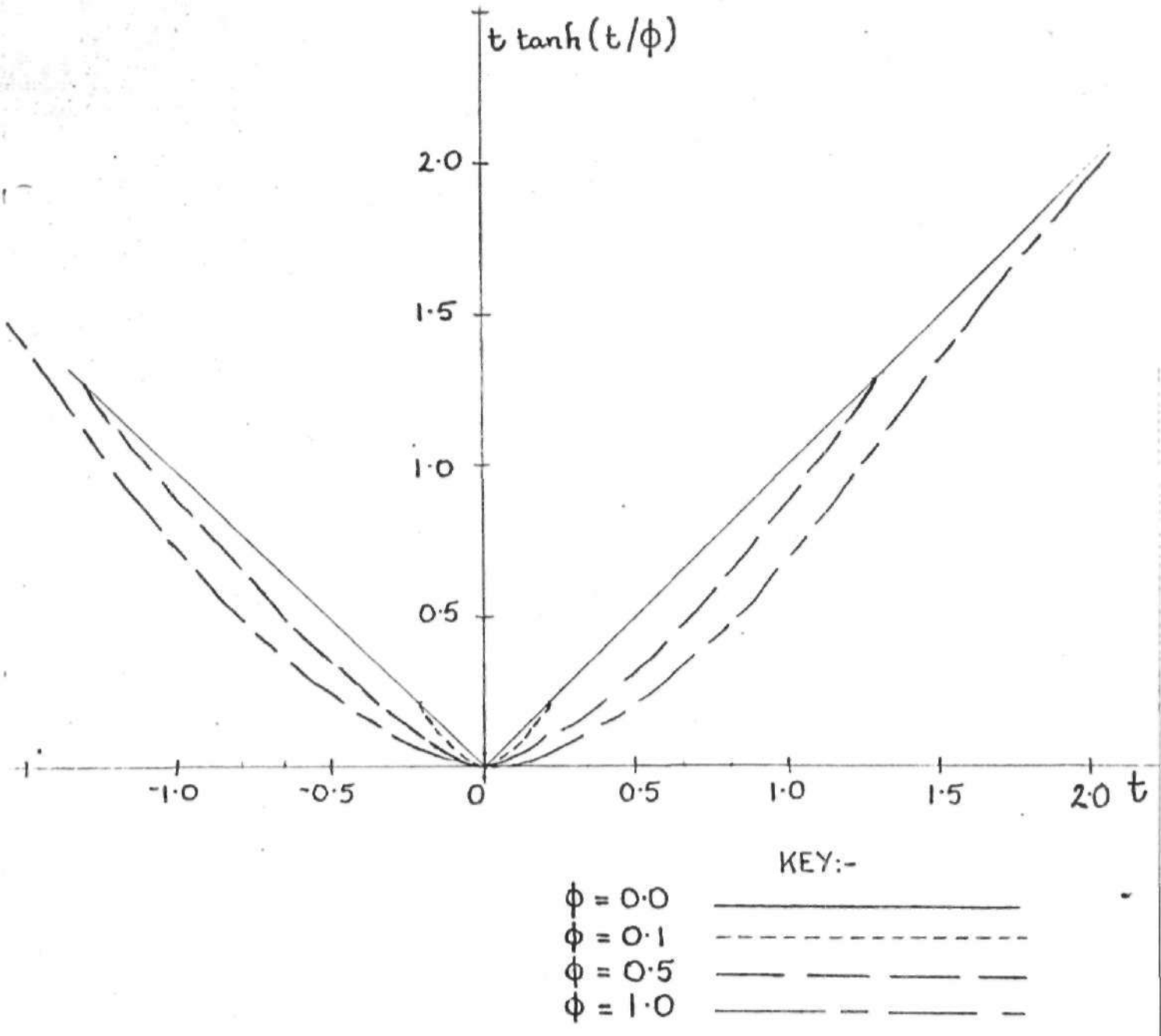


Figure 4.4.1

although the technique appears to be good in fitting a segmented regression model, Models (4.4.1) or (4.4.2), it can be argued, are fundamentally different to (2.2.3).

5. ESTIMATION ALGORITHMS FOR PIECEWISE REGRESSION MODELS

5.1 General Description

In Chapters 3 and 4 we discussed algorithms applicable to continuous models only. However, in this Chapter, we shall consider estimation procedures for discontinuous models - piecewise regression models. Some of the estimation procedures can also, with modifications, provide approximate solutions to segmented regression problems.

For piecewise regression models, the change-point from one model to another was referred to as a "jump-point" in Chapter 2. The estimates of jump-points occur in an interval. Suppose we consider n observations denoted by $(x_1, y_1)....(x_n, y_n)$ to which a two-segment piecewise linear regression model is to be fitted. Suppose the best line possible is fitted to the first subset given by $(x_1, y_1)....(x_i, y_i)$ and the second fitted to the remaining points $(x_{i+1}, y_{i+1})....(x_n, y_n)$. The parameter estimates for the segments are obtained by local least squares estimation procedures. It is clear that join-point estimation cannot be achieved here, since only γ in the interval $[x_i, x_{i+1})$ (in association with the same values of β_{11}, β_{12}, β_{21}, β_{22}) will give the same values of $E(Y_1)$, $E(Y_2)....E(Y_n)$ in general. This implies that in estimation in piecewise regression models, it is not practicable to distinguish between values of γ lying between the same two adjacent x-values. So, firstly, we shall be talking about possible interval values for γ. Secondly, it will be more convenient to talk of "Solution partitions" as an equivalent term for estimated change-points. The relationship between the two is that the first constitutes the set of equivalent estimates for the jump-point, while the second is the selected final solution.

The estimation procedures considered are based on various underlying principles. Some involve a combination of standard least squares

techniques and techniques related to clustering procedures, while others are based on the use of dynamic programming as a tool in order to obtain a solution partition. We shall first discuss procedures for solving two-linear segment piecewise regression problems, considering situations where the two segments have equal error variances, as well as cases of different segment variances. The two-linear segment solution methods will then be extended to deal with the more general case with more than two segments. We shall also describe some other estimation techniques used to obtain quick approximate solutions to the problem, especially when the data set is very large. Some of these results only provide solutions that yield local minima.

5.2 Estimation Technique for 2-Segment (equal variance) Piecewise Regression Model

Consider the n observations defined in Section 5.1. For the two-segment case with equal segment variances, it is a relatively straightforward application of the local least squares technique to fit a piecewise model to the data. Let RSS(1, i) and RSS(i+1, n) respectively be the residual sums of squares obtained for the first and second segments, when piecewise models are fitted to the first and second subsets of partition i. The procedure considers all possible partitions of the data into two disjoint subsets, still imposing the condition that each subset contains at least two points (minimum number of points needed to fit a line). For every partition, the estimates of the parameters of the segments are obtained by the local least squares estimation techniques defined in Section 3.2. Then the overall residual sum of squares is obtained by adding the residual sums of squares for the two subsets of the data; that is, for partition i,

we calculate

$$RSS(i) = RSS(1, i) + RSS(i+1, n).$$

The results are then tabulated in the form similar to that in Section 3.5. The results which give rise to the smallest overall residual sum of squares provide the solution partition and the estimates of the parameters for each segment. If the solution partition is partition i, the jump-point is estimated to lie in the interval $[x_i, x_{i+1}]$.

5.3 Extension to K>2 Segments (equal segment variance)

The estimation of a K-segment piecewise regression model problem may be performed by the use of dynamic programming. The concept as applied to piecewise regression models was illustrated by Guthery (1974) and Hawkins (1976). The approach discussed here is based on Hawkins' ideas based on the least squares principle.

Suppose we wish to fit a K-segment piecewise regression model to n data points. Dynamic programming uses some recurrence equation defined by

$$RSS^a(j,n) = \min_{j+1 \le S \le n-2(k-1)} \left[RSS^1(j,s) + RSS^{a-1}[S+1, n] \right]$$

where $RSS^a(j,n)$ is defined as the residual sum of squares when segments are fitted to the data points $(x_j, y_j) \ldots . (x_n, y_n)$.

Consider the application of $RSS^a(j,n)$; j should normally take the value 1, i.e. $RSS^a(1,n)$ to denote the residual sum of squares when segments are to be fitted to the n data points defined above. By similar reasoning, $RSS^{a-1}(j,s)$ denotes the residual sum of squares when (a-1) segments are fitted to the first (s-j+1) points, and $RSS^1(s+1, n)$ denotes the corresponding residual sum of squares when a single segment is fitted to the remaining (n-s) points.

For a clearer understanding of the technique, a 3-linear segment example is given below for illustration. Suppose that three linear segments are to be fitted to the data given in the schedule below, and

X	1.0	2.0	4.0	6.0	8.4	9.0	10.0	12.0
Y	5.0	4.0	2.8	4.2	5.0	5.1	2.8	2.3

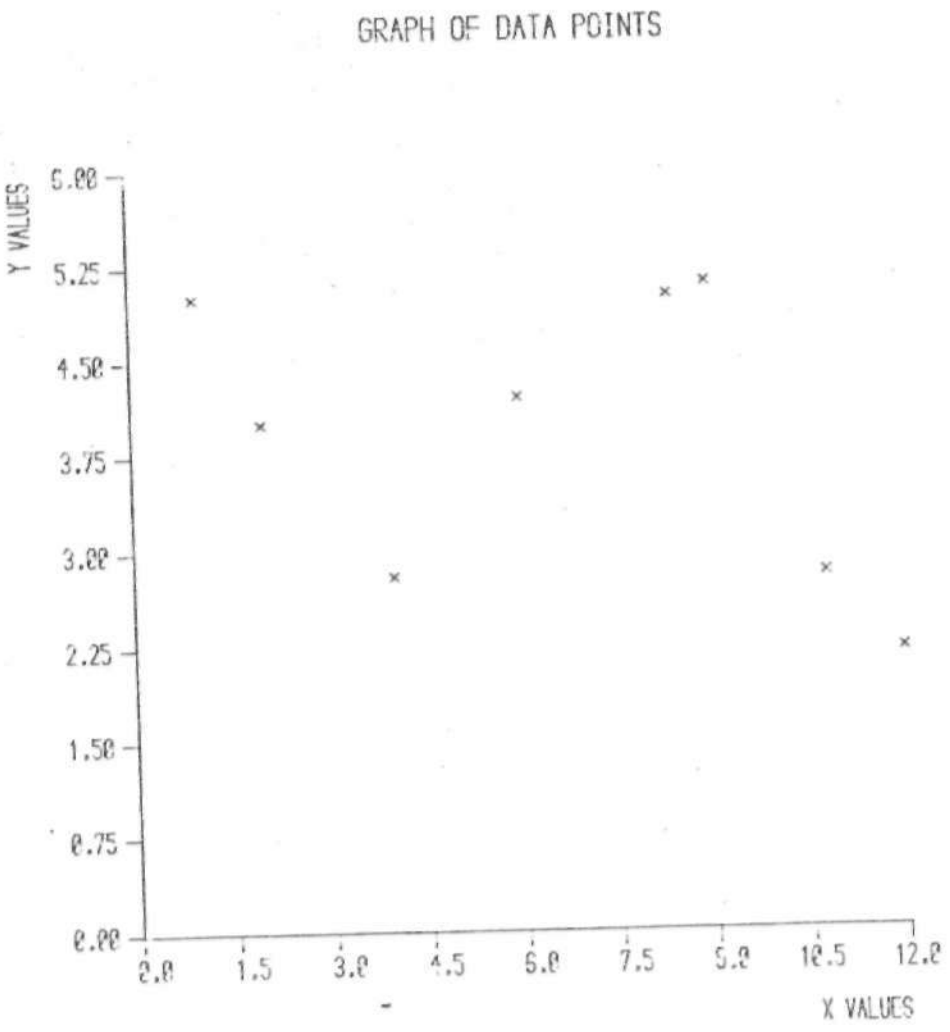


Figure 5.3

the results are given in Table 5.3 below.

Example

X	1.0	2.0	4.0	6.0	8.4	9.0	10.8	12.0
Y	5.0	4.0	2.8	4.2	5.0	5.1	2.8	2.3

The graph of the above data is given in Fig. 5.3.

Table 5.3

Table of $RSS^k(j, 8)$ a(No. of segments)

j	1	2	3
1			0.050
2			
3		0.153	
4		0.004	
5	0.332	0.000	
6	0.235		
7	0.000		

Now let us use the above results to explain how the procedure operates, as well as how the entries in the Table are obtained. The first stage considers all the data and proceeds to obtain a single segment, fitting data points 5 to 8, 6 to 8 and 7 to 8, thereby obtaining the entries in the first column, i.e.

$RSS^1(5, 8) = 0.332$

$RSS^1(6, 8) = 0.235$

$RSS^1(7, 8) = 0.000$ - only two points, hence there is a perfect fit here.

The second stage uses the recursion formula whose general form was given in Section 5.3. That is

$$RSS^2(j, 8) = \underset{j+1 \le S \le 6}{\text{Min}} \left[RSS^1(j, s) + RSS^1[s+1, 8] \right]$$

where i takes values $3 \le i \le 5$.

The above recursion is then used to obtain the best two-segment fits to data points 3 to 8, 4 to 8, and 5 to 8 yielding the following results:

$$RSS^2(3, 8) = \text{Min} \begin{pmatrix} RSS^1(3, 4) + RSS^1(5, 8) = 0.00 + 0.332 \\ RSS^1(3, 5) + RSS^1(6, 8) = 0.108 + 0.235 \\ RSS\ (3, 6) + RSS\ (7, 8) = 0.153 + 0.000 \end{pmatrix}$$

$$= 0.153,$$

$$RSS^2(4, 8) = \text{Min} \begin{pmatrix} RSS^1(4, 5) + RSS^1(6, 8) = 0.00 + 0.332 \\ RSS^1(4, 6) + RSS^1(7, 8) = 0.004 + 0.00 \end{pmatrix}$$

$$= 0.004;$$

$$RSS^2(5, 8) = RSS^1(5, 6) + RSS^1(7, 8) = 0.0 + 0.0$$

$$= 0.00.$$

The results appear in column two of the above table. Using the approach described in the previous section, consider the above results. The best two-segment fits given above are obtainable by choosing the solution partition resulting in the smallest residual sum of squares for all possible partitions of the given set of data points.

Finally, stage three uses the recursion

$$RSS^3(i, 8) = \underset{i+1 \le S \le 4}{\text{Min}} \left(RSS^1(i, s) + RSS^2(s+1, 8) \right)$$

and by similar argument leads to the following

$$RSS\ (1, 8) = \text{Min} \begin{pmatrix} RSS^1(1, 2) + RSS^2(3, 8) = 0.000 + 0.153 \\ RSS^1(1, 3) + RSS^2(4, 8) = 0.046 + 0.004 \\ RSS^1(1, 4) + RSS^2(5, 8) = 2.022 + 0.000 \end{pmatrix}$$

$$= 0.050.$$

For problems involving a larger number of segments, a table can be constructed such that it is possible to observe the minimizing value of S

for each entry. Such tables make it easier to see what the solution partition is. In column 3 of the above table, the value entered is the partition leading to the smallest residual sum of squares out of all possible three-segment partitions of data. The minimizing value of S is the corresponding value for the partition, i.e. S = 3. Hence the solution partition is given by

(1,2,3); (4,5,6); (7,8)

The estimated regression

$$E(Y) = 5.60 - 0.713x \qquad x<\gamma_1$$
$$= 2.35 + 0.310x \qquad \gamma_1 \leq x < \gamma_2$$
$$= 8.2 - 0.50x \qquad x \geq \gamma_2$$

$\gamma_1 \in (4.0, 6.0)$ and $\gamma_2 \in (9.0, 10.6)$.

Notice that the first stage fitted a single segment to points 5 to 8, 6 to 8, 7 to 8 only. This is due to the fact that we are restricting each segment to contain at least two points, to allow unique estimates for the slope parameter estimates. Hence, fitting single segments to points 4 to 8, 3 to 8 was not considered appropriate.

Dynamic programming has an important advantage as a tool for solving a K-segment piecewise regression problem, in that it does not require exhaustive consideration of all possible partitions of the given set of data. Hence some results only needed to be calculated once, and can be re-used later in the procedures (as can be seen in the recurrence equations in the above example). One noticeable problem, however, in the use of dynamic programming for large data sets, is the fact that, despite the above advantage, a prohibitively large number of subsets will be involved which could lead to heavy labour on the part of the user, or a large amount of computer time.

5.4 Estimation Technique for 2-Segment (unequal variance)

Quandt (1958) examined the problem when the segments have different variances, and also have normally distributed errors, using the maximum likelihood approach to obtain the estimates of the parameters. Suppose there are n data points to which two-segment piecewise models are to be fitted. Let the first segment be fitted to the points $(x_1, y_1)....(x_i, y_i)$ with parameters given by β_{11}, β_{12} and variances σ_1^2. Similarly, the second segment is fitted to the remaining points with parameters β_{21}, β_{22} and variance σ_2^2. The range where one segment is succeeded by another is denoted by an interval.

The procedure considers all possible partitions of the data into two subsets with the restriction that there are at least two points in each subset. For each partition i considered, the likelihood function is maximized with respect to the parameters given above, except for γ. This maximized value is, for any $\gamma \in (x_i, x_{i+1})$

$$L(i) = -n\log(2\pi) - i\log\hat{\sigma}_1(i) - (n-i)\log\hat{\sigma}_2(i) - \frac{n}{2} \qquad (5.4.1)$$

where $\sigma_1^2(i)$ and $\sigma_2^2(i)$ are the (biased) maximum likelihood estimates of the variances for the first and second segments obtained from local least squares. The procedure maximizes the function (5.4.1) over all values of i in the range $3 \le i \le n-3$ by calculating for each i the value of L(i). The value of i for which L(i) is maximized defines the solution partition of the data, and the corresponding parameter estimates constitute the final solution for the problem. The jump-point is then estimated to lie in the interval $[x_i, x_{i+1})$.

For the K-segment problem this approach is extended by means of dynamic programming discussed in Section 5.2. Hawkins (1976) used this approach to solve the general K-segment problems.

5.5 Other Estimation Techniques for Piecewise Regression Models with more than Two Segments

When the number of data points is large, the estimation procedures for estimating the parameters described in Section 5.3 become difficult because the number of possible partitions of the data into subsets is prohibitively large. This now leads to the introduction of a number of approximate methods. The techniques presented in the next two sections serve as quick methods of obtaining approximate solutions for the piecewise models.

5.5.1 The Search Algorithm

The algorithm was proposed by Ertel and Fowlkes (1976). It is designed mainly for Piecewise regression models, but the authors also used it to provide approximate solutions for segmented regression problems by restricting the join-point estimates to lie midway between two adjacent values of x. The search algorithm provides a "local minimum" solution. The procedure has the attractive feature that it is based on the standard least squares theory. The algorithm consists of two stages. The first consists of the choice of a starting partition which may be obtained from a plot of the data whereby the number of segments appropriate for the data is decided, and the approximate location of the change-points determined (the authors used a plot of standardized predicted errors in order to obtain their starting partition). The second stage involves iterative alteration of the partition in order to improve the overall residual sum of squares for the subsets.

For illustration, an example of a three-segment piecewise regression model will be considered. Suppose the starting partition obtained from the plot of data has three subsets containing n_1, n_2 and n_3 data points. Linear regressions are fitted to each of the subsets by local least squares estimation, and the overall residual sum of squares is obtained. A pertubation of the subset sizes, as below, is then considered. The

fitting is repeated using the subset sizes of the pertubation below with corresponding residual sum of squares being examined. The process stops when a local minimum is attained.

PERTUBATION	SUBSET SIZES
1	$(n_1; n_{2+1}; n_{3-1})$
2	$(n_1; n_2-1; n_{3+1})$
3	$(n_{1-1}; n_{2+1}; n_3)$
4	$(n_{1+1}; n_{2-1}; n_3)$

In order to achieve a fairly accurate solution, there is the need for a good choice of the starting partition, since this will enhance the pertubation to be conducted around the region where the change from one submodel to another actually occurs. Although the method does not provide a global minimum solution, it provides quick, approximate results.

5.5.2 Hierarchical Clustering Technique

The method was introduced by McGee and Carleton (1970) to fit K-segment Piecewise regression models. Hawkins (1976) also used a clustering technique for solving the same type of problem, but approached it from a different direction. Such algorithms are based on two theories - standard least squares theory and the ideas of standard clustering procedures. However, it must be noted that the term "hierarchical clustering" was used by the authors for convenience, because in the strict sense of the terminology, their technique is not based on similarity or dissimilarity matrices.

Hawkins' (1976) approach was based on a "divisive" technique in which all the data are initially regarded as a single group. The group is then successively split into smaller disjoint subsets. McGee and Carleton, on the other hand, adopted an agglomerative approach - the individual points initially constitute separate clusters which are built up into larger

subsets. The latter approach is described below. In the context of Piecewise regression the term cluster is used to denote the set of adjacent data points forming a subset, to which a line is to be fitted.

The technique does not require any graphical or other procedures to provide information about the number of segments in the data. The determination of K is based on the user's judgement in conjunction with the use of stopping rules, introduced to serve as guidelines. The description of the stopping rule and the other approaches for the determination of K, when unknown, are given in Section 5.6. The description of the algorithm is given for the general case.

Description of the Algorithm

We begin by making basic definitions which aid in the understanding of how the algorithm operates.

Minimal cluster This term denotes a cluster of the smallest allowable size to which a segment is to be fitted. In the case to be considered here with linear segments, the smallest allowable cluster consists of three adjacent points.

Potential cluster At any stage, the clusters being considered for formation are called Potential Clusters.

Fixed clusters The cluster chosen at any particular stage from the set of potential clusters becomes a fixed cluster, implying that the points in the cluster are associated with a line segment.

The algorithm proceeds by considering as a potential cluster all adjacent triples of points. A line is fitted to each and the corresponding residual mean squares are recorded. The potential cluster which leads to the smallest residual mean square is chosen as the first fixed cluster.

The next stage consists of three possibilities. Either

(I) the adjacent point to the left of the cluster is added; or

(II) the adjacent point to the right is added; or

(III) a new minimal cluster is formed among points that have not been clustered.

At either of the above stages I and II, the concept of goodness of fit is used. The goodness of fit of the points in a cluster is measured by the unbiased estimate of the error variance given by the residual mean square obtained.

All the possibilities are considered in the process and the appropriate residual mean squares are compared. Using a stopping rule described below, which serves as a guideline to the line of action necessary, the user decides on the basis of the rule. The stopping rule uses the F-distribution.

Consider the first two possibilities described above. Suppose a fixed cluster has been obtained at any stage and we wish to consider adding an adjacent point to the left or to the right of the fixed cluster. Let the point being considered be denoted by (x_r, y_r). Let the points in the fixed cluster be represented by $(x_{r+1}, y_{r+1}) \ldots (x_{r+t}, y_{r+t})$ containing t points. Suppose y_i denotes the fitted value for y_i based on the (t-r) points in the existing fixed cluster. Furthermore, suppose y_i^* denotes the fitted value for the point y_i under the estimation of the parameters based on (t-r+1) adjacent points. A statistic is calculated given by the ratio

$$\left\{ \left[y_r - \hat{y}_r^* \right]^2 + \sum_{i=r+1}^{t} \left[\hat{y}_i - \hat{y}_i^* \right]^2 \right\} \Big/ \phi$$

where ϕ is the residual mean square for the fixed cluster and is defined by

$$\phi = \sum_{i=r+1}^{t} [y_i - \hat{y}_i]^2 \; / \; (t-r-2)$$

The statistic is used to test whether a single linear regression model

applies to all the (t+1) adjacent points formed in the new cluster. Under the null hypothesis that a single linear regression applies, the distribution of the statistic is an F-distribution with degrees of freedom (1, t-r-2) provided the errors are normally, independently and identically distributed. The null hypothesis is rejected for large values of F and that point being considered is then not included in the new cluster.

Finally, the last possibility is the formation of a new cluster from the combination of two adjacent fixed clusters. For such situations, the stopping rule is based on the comparison between the appropriate residual mean squares under the null hypothesis that a single linear regression applies to the two clusters with the residual mean squares when a separate linear regression is fitted to each cluster. This statistic is given by

$$\frac{[RSS(j, t) - RSS(j, i) - RSS(i+1, t)]/2}{[RSS(j, i) + RSS(i+1, t)]/(t-j-3)}$$

where RSS(j, t) is the residual sum of squares obtained when a line is fitted to the combined cluster combining the points $(x_j, y_j)....(x_i, y_i)$ and $(x_{i+1}, y_{i+1})....(x_t, y_t)$. RSS(j, i) and RSS(i+1, t) are the residual sum of squares when separate lines are fitted to the above clusters. Under the null hypothesis stated above, the statistic given has an F-distribution with (2, t-j-3) degrees of freedom provided the condition stated earlier holds.

The two adjacent clusters are combined and a single line fitted for small values of the statistic. Otherwise, a single line is fitted to each cluster. It must be stated here that, because the stopping rules only serve as guidelines, the interpretation of the "significance level" must be done with caution. The process terminates when neither of the possibilities stated above is achieved. The authors provided no explicit termination rule for stopping the process. They remarked that the user is expected

to define a point at which the procedure is to be terminated. Because of the lack of the termination point, and differences in the choice of significance level, it has been known that various values of K (number of segments) have resulted.

To avoid this anomaly, it is recommended that the user incorporates the use of plotting procedures which enables the user to determine K approximately, as stated in Chapter 4. The authors did not make any recommendation on the use of plots of data. However, they suggested that several trial values for the minimal cluster sizes be considered and the process repeated for each value considered. They argued that the results should lead to the same solution.

We shall look at other methods recommended for the choice of approximate values for K in the next section.

One general comment about the effectiveness of Hierarchical procedure is that although it could serve as a useful tool for large data sets, there may be the shortcoming of having to deal with a prohibitively large number of potential clusters. This may however still be preferable to a full solution involving dynamic programming. Furthermore, there is, as already mentioned, the fact that there may be different solutions for various significance levels since the stopping rule depends largely on the choice made. Another point against its use is the fact that Hierarchical clustering procedures, as applied to Piecewise regression models, do not possess a strong theoretical basis.

5.6 Selection of the Number of Segments in Piecewise Regression Models

Most experiments in using either Segmented or Piecewise regression analysis make the assumption that either the number of segments is known or that it can be obtained from a plot of the data section. Some authors, however, make no mention of the use of graphs but suggest some other means

whereby the number of segments can be obtained. Some of these methods have no sound underlying theory, and so are best regarded as providing guidelines. Three of these methods are considered below. The first method is based on the use of an F-test as a stopping rule. This was discussed in the previous section and used by McGee and Carleton in their Hierarchical clustering technique. The method relates to the process of combining segments, such that two separate segments are represented by a single segment whenever the test of hypothesis, that a single line is applicable to the two separate segments, is not found significant for a chosen significance level, by calculating the relevant statistic discussed. The details and the shortcomings of such methods were given in the previous section.

The second method is based on the fact that the residual mean square tends to decrease with increasing K until the correct value of K is attained. Thereafter, it will be roughly constant. The user looks at the magnitude of the residual mean squares for various values of K and decides whether increasing K brings about a large decrease in the residual mean square to make the increase worthwhile. That is, one may look at the plot of RMS (residual mean squares) against K and choose the value of K where RMS "flattens out" (typically it may not necessarily be the minimum RMS).

A third method for choosing K is based on the principle underlying Mallows C_p-statistic in order to choose the value of K that best describes the data. Mallow's C_p-statistic was originally developed for selecting or choosing variables in a multiple regression analysis. The definition is given in general and later related to application in Piecewise regression models.

Definition Mallow's C_p-statistic is defined as

$C_p = RSS_p/\hat{\sigma}^2 - n+2p$, where p is the number of parameters to be

estimated in the model, RSS_p denotes the residual sum of squares for the model, based on n-p degrees of freedom, and $\hat{\sigma}^2$ denotes an estimate of the error variance σ^2 obtained by fitting a model with a suitably large number of segments. There is not much difficulty in applying the statistics to continuous models, since the situation is equivalent to the linear spline problems. However, with K-linear segment discontinuous models, p is chosen to be 2K (disregarding in this case the unknown jump-points). Ertel and Fowlkes remarked that, as the true value of K is approached, there is always a noticeable decrease in the corresponding statistic denoted by, say, $C_{(2K)}$. For a range of trial values of K, the corresponding statistics are calculated. Plotting these statistics against 2K the technique chooses a small value of K for which $C_{(2K)}$ is approximately equal to 2K.

The C_p-statistic has a number of advantages over McGee and Carleton's stopping rule guideline based on an F-test. The statistic possesses an established underlying principle and theory in regression analysis; however, not much is known of its application to Segmented or Piecewise regression models. This advantage is much less obvious since it is being applied to a situation which it has not been designed to deal with, and more so when the effect of the occurrences of the jump-points is ignored. Another advantage is that the statistic does not make any assumption about the normality and the independence of the errors as required by McGee and Carleton's F-test stopping rule.

Apart from the shortcoming mentioned above, the statistic has another shortcoming in that it is based on a vague concept of "approximately equal" for the choice of K to be determined. On the concept, it is possible to have more than one value of K satisfying the condition $C_{(2K)} \approx 2K$. For situations like that, it is recommended that the smaller

value of $C_{(2K)}$, for which the condition holds, be chosen. As a recommendation, it is advocated that graphical procedures be adopted, since they will help to clarify a number of points discussed above.

6. INFERENCES IN SEGMENTED AND PIECEWISE REGRESSION MODELS

Questions of inference in Segmented and Piecewise regression analysis have not been dealt with in the literature as fully as computational algorithms. Most of the work on inference has been based on the assumption that the number of segments in the model under consideration is known, occasionally with the additional assumption that the location of the change-points is known. The introduction of the second assumption allows the use of standard least squares distributional results.

Although inference is not just hypothesis testing alone, we shall consider briefly some of the areas a user might wish to employ in hypothesis testing in Segmented and Piecewise regression models. A brief discussion on the construction of confidence intervals will be given later in this Chapter.

On testing of hypothesis, firstly, the user may be interested in investigating the adequacy of, for instance, a one-segment regression against the alternative that there are two segments. Secondly, he might be interested in testing hypotheses about the change-points. Specifically in segmented regression models, the experimenter frequently wishes to test the hypothesis that the join-points occur at some predetermined values, say, $\underline{\gamma} = \underline{\gamma}_o$. Such questions arise, for example, if one wishes to test the hypothesis that the optimum rate of growth is achieved at a particular level of, say, a given nutrient in a fertilizer experiment or in a poultry drug testing experiment on the farm. The same problem could be approached *via* the construction of confidence regions for the estimated value of $\underline{\gamma}$. Similarly, confidence regions may be constructed for the slope parameters, or differences between slope parameters, provided the appropriate asymptotic results hold.

Quandt (1961) derived a likelihood ratio test statistic for a problem in which the error variances for the two segments are not assumed

equal. He examined the null hypothesis that a single-segment will fit as adequately as a two-segment discontinuous linear regression model. He investigated the properties of the test using simulated data, and found that the conjecture that the likelihood ratio test statistic $-2\log_e\lambda$ is distributed, under the null hypothesis of a single-segment fit, approximately as χ^2 with 4 degrees of freedom, was rejected for each set of his simulated data, at a significance level of 0.00001. He concluded that the χ^2 approximation had failed completely. His argument for $\chi^2(4)$ is based on the difference between the number of parameters under the null and alternative hypotheses. Under the null hypothesis there are three parameters, viz.

$$\beta_{11} = \beta_{12} = \beta_1$$

$$\beta_{21} = \beta_{22} = \beta_2$$

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

that is, the parameters are equal for the two segments under the null hypothesis, implying a single-segment fit to the data. Under the alternative hypothesis, that there are two segments, there are seven parameters $(\beta_{11}, \beta_{12}, \sigma_1^2, \gamma, \beta_{21}, \beta_{22}, \sigma_2^2)$. On the basis of the standard argument, one expects $-2\log_e\lambda$ to have a χ^2 distribution with (7-3) degrees of freedom. Describing the nature of the failure Quandt observed that, in the empirical distribution of $-2\log_e\lambda$, observations tended to have larger values than the corresponding values for the $\chi^2(4)$ distribution. His theoretical explanation for the failure was based on the fact that the test statistic has a χ^2 distribution only if the derivative of the likelihood function disappears at a maximum, and the maximum likelihood estimators are approximately normally distributed. He remarked that the assumption may not be valid for a discontinuous function of γ for any n (total number of data). In fact for continuous models, the likelihood function is a continuous function of γ, but standard asymptotic results still do

not hold. As a way of overcoming the anomaly, Quandt suggested the use of small sample test statistics based on non-parametric tests by Kruskal-Wallis (1952) and Kiefer (1959). As a final alternative hypothesis Quandt recommended the use of a standard F-test for testing the null hypothesis (comparison of the variances under the null and the alternative hypothesis). Sprent tested the null hypothesis of one-segment fit against a two-segment alternative, using an F-test by comparing the residual mean squares under the null and the alternative. He also developed a likelihood ratio test statistic for the testing of the hypothesis, that the join-point takes specific value $\gamma = \gamma_0$.

Hinkley (1969, 1971) studied the problem of estimating and making inferences about the join-point for a continuous two-linear segment model. He derived the asymptotic distribution of the maximum likelihood estimates of the parameters in two segments, including that of the join-point, as well as the construction of confidence regions for the parameters using the asymptotic results.

There are arguments against the use of the asymptotic results for the construction of, for instance, the confidence regions. Firstly, the variances depend on the unknown parameters, in particular the variance of $\hat{\gamma}$, which as Hinkley remarked depends on the difference between the two slope parameters. The second argument is based on the fact that the asymptotic distributions may be poor approximations since the estimates are no longer asymptotically, normally distributed. Hinkley argued that for moderate sample sizes the asymptotic distributions of $\hat{\beta}_{12}$, $\hat{\beta}_{22}$ and $\hat{\gamma}_1$ are approximated by their asymptotic normal forms. He also discussed further problems associated with hypothesis testing relating to γ. He examined the properties of the distribution of the parameter estimates by simulation.

On hypothesis testing, Hinkley used the likelihood ratio test

statistic to test the hypothesis of a single-segment model, against the alternative of two segments, using a modified form of the type of two-segment model given in Chapter 2, where the model involved has a continuity constraint already imposed on the model. He observed that under the null hypothesis, the distribution of the test statistics $-2\log_e\lambda$, was approximately χ^2 with 3 degrees of freedom.

Standard asymptotic theory results of likelihood ratio tests would lead one to expect a $\chi^2(2)$ using the following argument. The null hypothesis model has parameter β_1, β_2 and σ^2 (one segment), while the alternative hypothesis of two-segments fit has the parameters β_1, β_{12}, β_{22}, γ and σ^2 (using Hinkley's form of model in which β_1 represents $E(y/\gamma)$ leading to χ^2 distribution with (5-3) degrees of freedom). The explanation tion was that the rejection is connected with the fact that the estimates have an asymptotic distribution and that they are not normally distributed as anticipated earlier. In particular, $\hat{\beta}_{22} - \hat{\beta}_{12}$ and γ have non-normal asymptotic distributions when $\beta_{22} = \beta_{12}$.

For a similar test, Feder (1975a) showed that under the null hypothesis the asymptotic distributions of $-2\log_e\lambda$ is that of the maximum of a number of correlated $\chi^2(1)$ and $\chi^2(2)$ random variables, varying with the configurations of the observed values of the independent variable. Feder (1975b) investigated the properties of the asymptotic distributions of the estimates geometrically, and came up with a suggestion that, instead of using the asymptotic estimates that have non-normal distributions, the asymptotic chi-square results of Wilks (1938) can be used in order to perform the above test of hypothesis. Feder also found out that in general, when there are fewer segments than assumed, the least squares estimates are not asymptotically normal, and the log-likelihood ratio test statistic, for testing a single-segment fit against an alternative of two segments, is no longer asymptotically χ^2 distributed. Finally, from the illustration

of the test of null single-segment fit, against the alternative of a two-segment fit, it can be stated that there is no easy way of testing the hypothesis for the general K-segment problem, since the standard results can only be used when the null hypothesis is not true.

7. SUMMARY AND CONCLUSION

The problem of fitting linear Segmented or Piecewise regression models is not an easy one where more than two segments are involved. However, for most practical situations, one is usually confronted with no more than two-segment problems, except in a few odd situations. Hence for situations with only two segments, already established algorithms, as well as the proposed one, are suitable for solving the problem. They have the features that make them fairly automatic, and they do not require much graphical representation of the data, other than for establishing the fact that Segmented or Piecewise regression will be the most suitable model to adopt, rather than getting involved with complicated, polynomial functions.

For more than two-segment problems, it is always helpful and highly advisable that the experimenter plots his data in order to establish the optimum number of segments that best describes his data, and to choose between fitting Piecewise or Segmented regression models. For most practical situations, where information is required of the location of change-points in the system, it is evident that the Segmented regression models will be more appropriate and have features which have practical interpretation, as seen in the section on application.

Piecewise regression models are only useful when no information is required of the change-points, or when they have no practical interpretation.

Finally, in order to fit either Segmented or Piecewise regression models, the objectives of the experiment must be clearly studied; and the adequacy of such models considered, by plotting the data, before embarking on fitting either a Piecewise or a Segmented regression model.

Since most of the examples of data sets were constructed artificially,

much consideration has been on the estimation, rather than on the inferences. Thus most of the work presented is a brief discussion of what has been achieved in this area without reproducing rigorous mathematical results.

REFERENCES

1. Bacon, D.W. and Watts, D.G., Estimating the transition between two regression systems obeying two separate regimes, Biometrika, 58 (1971), pp.525-534.

2. Beckman, R.J. and Cook, R.D., Testing for two phase regressions, Technometrics, 21 (1979), pp.65-69.

3. Bellman, R. and Roth, R., Curve fitting by segmented straight lines, J. Am. Stat. Assoc., 64 (1969), pp.1079-1084.

4. Ertel, J.E. and Fowlkes, E.B., Some algorithms for linear spline and Piecewise multiple linear regression, J. Am. Stat. Assoc., 71 (1976), pp.640-648.

5. Feder, P.I., On asymptotic distribution theory in segmented regression problems - Identified case, Annals of Statistics, 3 (1975a), pp.49-83.

6. Feder, P.I., The Log-likelihood ratio in segmented regression, Annals of Statistics, 3 (1975b),pp.84

7. Fuller, W.A., Grafted Polynomials as approximating functions, Australian Journal of Agric. Econ., 13 (1969), pp.35-46.

8. Fisher et al., A model for the description and prediction of the response of laying hens to amino acid intake, Br. Poultry Science, 14 (1973), pp.469-484.

9. Hawkins, D.M., Point Estimation of the parameters of Piecewise regression models, Applied Statistics, 25 (1976), pp.51-57.

10. Hinkley, D.V., Inferences about the intersection in two phase regression, Biometrika, 56 (1969), pp.495-504.

11. Hinkley, D.V., Inferences in two phase regression, J. Am. Stat. Assoc., 66 (1971), pp.736-742.

12. Hudson, D.J., Fitting segmented curves whose join points have to be estimated, J. Am. Stat. Assoc., 61 (1966), pp.1097-1129.

APPENDIX

CONSTRUCTION OF CONSTRAINED LEAST SQUARES ESTIMATES FOR THE GENERAL LINEAR MODEL (FULL RANK CASE ONLY)

Consider a full rank linear model defined by

$$\underline{Y} = \underline{X\beta} + \underline{\varepsilon} \tag{1}$$

Where $\underline{Y}$ is an (nxl) vector of responses, $\underline{X}$ is an (nxp) design matrix of rank p and $\underline{\beta}$ is a (pxl) vector of parameters. $\underline{\varepsilon}$ is the vector of error terms assumed to be independently and normally distributed with mean zero and variance σ^2.

The sum of squares

$$S = \|\underline{Y} - \underline{X\beta}\|^2 \tag{2}$$

is to be minimized subject to the constraint

$$\underline{K\beta} = \underline{b} \tag{3}$$

where $\underline{K}$ is (qxp) matrix of rank q.

Using the theory of Lagrange multipliers, we seek to minimize

$$S = \|\underline{Y} - \underline{X\beta}\|^2 + \underline{\lambda}' (\underline{K\beta} - \underline{b}) \tag{4}$$

where $\underline{\lambda}$ is a vector of Lagrange multipliers.

Differentiating S with respect to $\underline{\beta}$ and $\underline{\lambda}$, we obtain

$$\frac{\partial S}{\partial \beta} = -2\underline{X}'\underline{Y} + 2\underline{X}'\underline{X\beta} + \underline{K}'\underline{\lambda} \tag{5}$$

$$\frac{\partial S}{\partial \lambda} = \underline{K\beta} - \underline{b} \tag{6}$$

Setting (5) equal to zero we have, for the constrained estimate denoted by $\hat{\beta}_c$

$$(\underline{X}'\underline{X})\underline{\hat{\beta}}_c = \underline{X}'\underline{Y} - \tfrac{1}{2}\underline{K}'\underline{\lambda}$$

and thus

$$\underline{\hat{\beta}}_c = (\underline{X}'\underline{X})^{-1}\,\underline{X}'\underline{Y} - \tfrac{1}{2}(\underline{X}'\underline{X})^{-1}\,\underline{K}^1\underline{\lambda} \tag{7}$$

Setting (6) equal to zero and substituting for $\underline{\hat{\beta}}_c$ from (7) we obtain

$$\underline{b} = \underline{K}\,(\underline{X}'\underline{X})^{-1}\,\underline{X}'\underline{Y} - \tfrac{1}{2}\underline{K}(\underline{X}'\underline{X})^{-1}\,\underline{K}'\underline{\lambda} \tag{8}$$

From the fact that

$$\underline{K}(\underline{X}'\underline{X})^{-1}\,\underline{K}' = |\underline{K}'\underline{K}| \;/\; |\underline{X}'\underline{X}| > 0$$

we see that

$\underline{K}(\underline{X}'\underline{X})^{-1}\,\underline{K}'$ is non-singular.

Thus, solving for $\hat{\lambda}$ in (8), we have

$$-\tfrac{1}{2}\hat{\lambda} = [\underline{K}(\underline{X}'\underline{X})^{-1}\,\underline{K}']^{-1}\,[\underline{K}(\underline{X}'\underline{X})^{-1}\,\underline{X}'\underline{Y} - \underline{b}] \qquad (9)$$

substituting for $-\tfrac{1}{2}\hat{\lambda}$ in (7) above, we have

$$\hat{\underline{\beta}}_c = (\underline{X}'\underline{X})^{-1}\,\underline{X}'\underline{Y} - (\underline{X}'\underline{X})^{-1}\,\underline{K}'[\underline{K}(\underline{X}'\underline{X})^{-1}\,\underline{K}']^{-1}[\underline{K}(\underline{X}'\underline{X})^{-1}\,\underline{X}'\underline{Y} - \underline{b}]$$

$$= \hat{\underline{\beta}} - (\underline{X}'\underline{X})^{-1}\,\underline{K}'\,[\underline{K}(\underline{X}'\underline{X})^{-1}\,\underline{K}']^{-1}\,[\underline{K}\hat{\underline{\beta}} - \underline{b}] \qquad (10)$$

where $\hat{\underline{\beta}}$ denotes the vector of unconstrained least squares estimates.

The residual sum of squares is given by

$$RSS_c = \|\underline{Y} - \underline{X}\hat{\underline{\beta}}_c\|^2$$

$$= \|\underline{Y} - \underline{X}\hat{\underline{\beta}} + \underline{X}\hat{\underline{\beta}} - \underline{X}\hat{\underline{\beta}}_c\|^2$$

$$= \|\underline{Y} - \underline{X}\hat{\underline{\beta}}\|^2 + \|\underline{X}\hat{\underline{\beta}} - \underline{X}\hat{\underline{\beta}}_c\|^2 + \text{cross product terms.}$$

The cross product term is

$2(\underline{Y} - \underline{X}\hat{\underline{\beta}})'\;(\underline{X}\hat{\underline{\beta}} - \underline{X}\hat{\underline{\beta}}_c)$ which disappears

since

$$\underline{X}'(\underline{Y} - \underline{X}\hat{\underline{\beta}}) = \underline{0}$$

This leaves us, after further simplification, with

$$RSS_c = \|\underline{Y} - \underline{X}\hat{\underline{\beta}}\|^2 + \|\underline{X}(\underline{X}'\underline{X})^{-1}\,\underline{K}'[\underline{K}(\underline{X}'\underline{X})^{-1}\,\underline{K}\,]^{-1}\,[\underline{K}\hat{\underline{\beta}} - \underline{b}]\|^2$$

$$= \|\underline{Y} - \underline{X}\hat{\underline{\beta}}\|^2 + [\underline{K}\hat{\underline{\beta}} - \underline{b}]^1\,[\underline{K}(\underline{X}'\underline{X})^{-1}\,\underline{K}']^{-1}\,[\underline{K}\hat{\underline{\beta}} - \underline{b}] \qquad (11)$$

The first term in (11) is the usual error sum of squares in unconstrained least squares estimation while the second term corresponds to the extra sum of squares due to the imposed constraint.