EFFECTS OF MEASUREMENT ERROR IN PIECEWISE REGRESSION MODELS

by

Rebecca Ann Teeter

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1400

April 1982

EFFECTS OF MEASUREMENT ERROR IN PIECEWISE REGRESSION MODELS

by

Rebecca Ann Teeter

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill

1982

Approved by:

Advisor

Reader

Reader

Reader

ABSTRACT

REBECCA ANN TEETER. Effects of Measurement Error in Piecewise Regression Models. (Under the direction of C.M. SUCHINDRAN.)

A study is made of the join point estimate in a piecewise linear regression model when measurement error is present in the independent variable. Initially, the effects of measurement error on a standard estimation technique for piecewise regression are investigated. Later, a new estimation technique is proposed which is not affected by measurement error.

The effects of measurement error on the ordinary least squares (OLS) estimate of the join point, as proposed by Hudson (1966), are examined by obtaining the asymptotic expected value and mean squared error (MSE) using a Taylor series approximation. Furthermore, the empirical distribution of the join point estimate is found from a simulation study. Comparisons are then made between the asymptotic and empirical results for various values of measurement error variance. It is found that the join point estimate based on OLS estimation is biased and the MSE increases with increasing measurement error.

Following Wald (1940), an estimation procedure is proposed which is based on a grouping technique known to be asymptotically unbiased for the regression coefficients. In addition to a simulation study parallel to that done for the OLS estimate, a numerical technique is introduced which enables one to obtain an estimate of the variance of the join point estimate from the sample data. Results from this variance estimate are compared with the variance from the empirical distribution obtained through the simulation study. The grouping approach is found to yield

consistent join point estimates and the numerical estimate of the variance is close to the variance from the empirical distribution.

A less comprehensive study is made of estimation procedures as proposed by Tukey (1951) which can be used when the data are replicated. In this case it is possible to obtain estimates of the error variances and use this additional information in estimating the regression coefficients and join point of the piecewise model.

The techniques are applied to the problem of estimating day of ovulation based on a woman's basal body temperature taken throughout a menstrual cycle.

ACKNOWLEDGMENTS

I would like to express my appreciation to my advisor, Dr. C.M. Suchindran, for his patience, guidance, and advice during the course of this research. I am also indebted to the other committee members, Drs. James Abernathy, Karl Bauman, C.E. Davis, B.V. Shah and John Taulbee for their helpful comments and suggestions. I especially thank Dr. B.V. Shah for his helpful suggestions and ideas that went into this dissertation.

In addition, special thanks go to Dr. Shrikant I. Bangdiwala whose support and encouragement have meant a great deal to me. I wish also to thank my parents, Carroll and Ann Teeter, as well as my siblings, Alice, Ellen and David, for their encouragement and understanding throughout my life, but especially in my academic pursuits. Special thanks also go to Drs. George and Jane Webb and Drs. Richard and Judith Brown and their families for their guidance, support and understanding throughout the pursuit of this degree.

I want also to acknowledge my support group which included Drs. Mary Foulkes, Diane Makuc, and Alcinda Lewis. Their continued support and encouragement even after they had finished and left to pursue their own careers has been greatly appreciated.

Many thanks go to the staff and students of the Department of Biostatistics who have helped and encouraged me throughout my stay here.

Finally, I would like to acknowledge the skillful typing of the manuscript by Ms. Stacy Miller and Ms. Jean McKinney and the financial support provided by the Department of Biostatistics and the National Institute of Child Health and Human Development grants No. HD00371 and No. HD07102 that enabled me to pursue this research.

RAT

TABLE OF CONTENTS

Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION AND REVIEW OF THE LITERATURE

1.1 INTRODUCTION

When one is dealing with time or another variable which imposes an order on the data one may wish to estimate the regression function by a series of submodels which join at their end points. Thus, one has a model which is composed of several segments rather than the single continuous function frequently used in regression problems. The points of intersection of the submodels are referred to as join points. Examples of such situations are found in many different disciplines. An example, from the biological sciences, is given by Mellits (1968) in his approach to the study of human growth. In one model total body water (TBW) is regarded as a linear function of height and weight. An investigation of height alone indicated that height had a non-linear effect on TBW. Thus, the final model which was fitted consists of two linear submodels which differ according to the range of the variable height. A second example regresses basal body temperature on the day of a woman's menstrual cycle. In this case one submodel would fit prior to the day of phase shift and a second submodel would fit after ovulation. Here, the join point or day of ovulation is unknown and, therefore, must be estimated as a parameter of the model. In all these examples the join points have a physical interpretation in that they correspond to some structural change in the underlying model. The fitting of submodels is generally referred to as piecewise linear regression (PWLR).

In regression problems one assumes that the measured Y values vary from their true values only by a random error component and one can assume that the X values are known constants or that X is a random variable whose distribution does not involve the regression parameters. With either of these assumptions one is able to use least squares procedures to obtain estimators which are unbiased and have minimum variances.

However, if X is measured with error or if Y has a component of error which is not random then parameter estimates must be adjusted accordingly. For example, measurements of height and weight in the first example may be in error if the devices used to take these measurements are inaccurate or the person using the devices consistently takes biased readings.

The focus of this research is a study of how various types of error in the variables affect the estimation of model parameters in the case of piecewise regression. Of particular interest is the effect of error in the variables on the estimation of the join point.

Current piecewise regression techniques are reviewed in section 1.2 . Section 1.3 contains a review of various procedures available to adjust for errors in measurement. Finally, section 1.4 outlines the proposal for research conducted in this thesis.

1.2 PIECEWISE LINEAR REGRESSION (PWLR)

1.2.1 Models

The fitting of submodels is generally referred to as piecewise regression. Other names arise depending on the types of functions included as submodels and the characteristics of these functions at their join points. The degree of difficulty encountered in fitting a piecewise model depends, in part, on how much knowledge one has concerning the number of submodels and the placement of the join points. In general, a piecewise regression model of r segments has the form:

$$E(y) = \begin{cases} f_1(\underset{\sim}{x};\underset{\sim}{\beta}_1) & \text{for } A_o < x_i < C_1 \quad \text{for some } i \\ f_2(\underset{\sim}{x};\underset{\sim}{\beta}_2) & C_1 < x_i < C_2 \\ \vdots & \\ f_r(\underset{\sim}{x};\underset{\sim}{\beta}_r) & C_{r-1} < x_i < A_r \end{cases}$$

where A_o and A_r are given constants. The C_j can be known or unknown. In the latter case the C_j are estimated as parameters in the model along with the $\underset{\sim}{\beta}_j$. The submodel $f_j(\underset{\sim}{x};\underset{\sim}{\beta}_j)$ is usually linear in $\underset{\sim}{\beta}_j$ and differentiable with respect to x. The most frequent examples are polynomials in x of degree q with coefficients $\underset{\sim}{\beta}_j$. Other forms are possible for the submodel such as those which incorporate trigonometric functions of x. In general, it is not necessary for $f_j(\underset{\sim}{x};\underset{\sim}{\beta}_j)$ to be of the same form for all j. However, there is one important constraint on these models. The submodels must join at their endpoints, $x_i = C_j$, to form a continuous overall model. If the C_j are known this constraint is simply a linear constraint on the unknown $\underset{\sim}{\beta}_j$ and $\underset{\sim}{\beta}_{j+1}$. However, when C_j is unknown the constraint becomes a non-linear one on C_j, $\underset{\sim}{\beta}_j$ and $\underset{\sim}{\beta}_{j+1}$. Solutions are more difficult to obtain for the latter case since the usual least squares estimators may not always be appropriate. Further restrictions can be imposed upon the model such as those constraints found when fitting spline functions.

A spline function is defined by Smith (1979) as a piecewise polynomial of degree n whose function values agree at the points where they join. In addition, the first n-1 derivatives also agree at the join points. In spline theory, much of which is adopted from the field of engineering, the join points are referred to as knots. Since fewer than the maximum number of continuity restrictions are met by the piecewise models defined earlier this piecewise model could be viewed as a special case of a spline function and the terms are often confused in the literature. It is also possible to impose more restrictions on spline functions such as a requirement that the second derivatives be positive. Such additional restrictions make the fitting of models an even more difficult non-linear problem.

Feder (1975) makes another important distinction between piecewise models and splines. He points out that in spline approximation the knots are chosen merely for analytical convenience while in piecewise regression theory the changeover points between segments have intrinsic physical meaning in that they correspond to structural changes in the underlying model. Granted that both piecewise regression and the spline technique share many of the same technical problems, this review will deal only with piecewise models as defined herein and no attempt will be made to draw on all the theory and literature dealing with splines.

Specifically, consider the following model.

Let $\underset{\sim}{Y}$ be an n-dimensional response vector, $\underset{\sim}{X}$ an n by p dimensional matrix containing a column of 1's and p-1 independent variables, and β_k a p by 1 vector of regression coefficients. The general linear regression model can then be written:

$$f(\underset{\sim}{X}) = E(\underset{\sim}{Y}) = \underset{\sim}{X}\underset{\sim}{\beta}_k \tag{1.1}$$

$E(\underset{\sim}{Y})$ can be viewed as a p-dimensional response surface. This response surface can be cut into two sections by cutting the domain of the response surface with a p-1 dimensional surface. Let $g(\underset{\sim}{X}^*) = C_k$ be some function of the p-1 independent variables, represented by $\underset{\sim}{X}^*$, which defines such a cut. The general PWLR model with r sections can then be written:

$$f(\underset{\sim}{X}) = E(\underset{\sim}{Y}) = \underset{\sim}{X}\underset{\sim}{\beta}_k = f_k(\underset{\sim}{X};\underset{\sim}{\beta}_k) \tag{1.2}$$

$$\text{for } C_{k-1} \leq g(\underset{\sim}{X}^*) \leq C_k \quad ; \quad k = 1,2, \ldots ,r$$

with the restriction that the functions f_k and f_{k+1} are equal at their join point; that is, they are equal where they meet on the surface $g(\underset{\sim}{X}^*) = C_k$. I.e.,

$$f_k(\underset{\sim}{X}^*;\underset{\sim}{\beta}_k) = f_{k+1}(\underset{\sim}{X}^*;\underset{\sim}{\beta}_{k+1}) \tag{1.3}$$

The methods of estimating the parameters in such a PWLR model can be divided into two cases depending on whether the value of the join point is known in advance or must be estimated from the data.

1.2.2 Known Join Points

If the abscissae of the join points are known, the constraint (1.3) is simply a linear constraint on the unknown parameters β_k and β_{k+1}. Consider the two dimensional case with model:

$$E(y_i) = \begin{cases} \beta_{10} + \beta_{11}x_i & \text{when } A_0 \le x_i \le C_1 \\ \beta_{20} + \beta_{21}X_i & C_1 \le x_i \le C_2 \\ \vdots & \vdots \\ \beta_{r0} + \beta_{r1}x_i & C_{r-1} \le x_i \le A_r \end{cases} \tag{1.4}$$

and with the constraint:

$$\beta_{k0} + \beta_{k1}C_k = \beta_{k+1\ 0} + \beta_{k+1\ 1}C_k \ ; \quad k = 1,2,\ldots,r-1 \tag{1.5}$$

In this situation solutions for the unknown parameters β_k and β_{k+1} can be found very easily with the aid of indicator variables. The model (4) can be reparameterized and written:

$$E(y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2(X_{i1}-C_1)X_{i2} + \beta_3(X_{i1}-C_2)X_{i3}$$

$$+ \ldots + \beta_{k+1}(X_{i1}-C_{r-1})X_{ik+1} \tag{1.6}$$

where

X_{i1} = independent variable

$$X_{ij} = \begin{cases} 0 \text{ when } X_{i1} \le C_j \\ 1 \qquad\quad X_{i1} > C_j \end{cases} \qquad j = 2,3,\ldots,r-1 \tag{1.7}$$

With this reparameterization one can now use ordinary regression techniques to fit the model.

Monti, Koch, Sawyer (1978) employ this type of model with three segments and a slightly different parameterization to study rat thyroid growth over time in a cross-sectional growth experiment. Results from this and previous studies indicated that rapid growth occurred during the first twelve days or first phase. The growth was less rapid during the second phase, and during the final phase a plateau-like state was reached in which little growth occurred. The piecewise regression technique is advantageous in a case such as this because it enables one to take into account distinct characteristics (e.g. growth rates) of each phase while maintaining the continuity of the model across the phase. In addition, the linear nature of the final segment provides a basis for predicted values which are not overly sensitive to variation in the pattern of growth in earlier phases. A polynomial regression model would not be appropriate since polynomials would not provide a satisfactory representation of the data in the final phase where little growth occurred. It is the objective of the experimenter to compare four groups of rats at the end of the study rather than provide a growth model for rat thyroids throughout the entire experiment. Hence, the segmented linear regression model provides a simple and easily interpreted model which is descriptively useful and reflects group differences in the final phase.

Imposing additional constraints on the model, Fuller (1969) fits what he refers to as "grafted polynomials". He first presents a method for fitting a second order polynomial with the additional constraint that the first derivatives be equal at the join points. Next, he extends his model to the three dimensional case where the domain of the response surface is a plane defined by the values of the variables X_1 and X_2. A three-dimensional model with two sections can be written:

$$E(\underset{\sim}{y}) = \begin{cases} \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_1^2 + \beta_{14}X_2^2 + \beta_{15}X_1X_2 \\ \\ \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_1^2 + \beta_{24}X_2^2 + \beta_{25}X_1X_2 \end{cases} \quad (1.8)$$

for $X_1 \leq K_0 + K_1X_2,\ K_1 \neq 0;\quad X_1 > K_0 + K_1X_2$

In this model the line $X_1 = K_0 + K_1X_2,\ K_1 \neq 0$ divides the (X_1,X_2) plane into two sections or half-planes and the model which differs in each of these subdomains is constrained to join on the line $X_1 = K_0 + K_1X_2$. The additional constraints imposed on this model are that the partial derivatives with respect to X_1 are equal as well as the partial derivatives with respect to X_2. These constraints allow for the direct estimation of 6 parameters -- β_{10}, β_{11}, β_{12}, β_{13}, β_{14}, $\beta_{15} + (\beta_{25}-\beta_{15})$. The remaining parameters are defined as functions of the 6 estimated parameters. With these parameter definitions the reparameterized model is then written as a single regression equation which resembles the first equation of the original parameterization plus a variable Z with coefficient $(\beta_{25}-\beta_{15})$. Here Z is defined as:

$$Z = \begin{cases} 0, & X_1 \leq K_0 + K_1X_2;\ K_0 \neq 0 \\ \\ -(2K_1)^{-1}(X_1 - K_0 - K_1X_2)^2, & X_1 > K_0 + K_1X_2 \end{cases} \quad (1.9)$$

As a further step Fuller discusses parameterizations used when constraints that higher-order derivatives of the response surface be equal at the join are placed on the model.

Finally Fuller includes a discussion of the use of grafted polynomials to approximate trends in a time series. He points out, as did

Monti et al., that fitting piecewise models with a linear final segment is, in many cases, superior to fitting a polynomial, especially when one is interested in prediction or extrapolation of the series. By placing constraints on the derivatives at the joins such that the linear trend is continuous and tangent to the non-linear trend one can still obtain a smooth curve which does not go to infinity as rapidly as a higher order polynomial.

1.2.3 Unknown Join Points

Cases in which one is fitting a PWLR model when the abscissae of the join points are unknown present a more complex estimation problem than those cases in which the join points are known since the continuity constraint is nonlinear in the unknown parameters -- β_k, β_{k+1} and C_k. Consider again the two dimensional case with r segments. The model has the same form as (1.4) with constraint (1.5); however, it is obvious that the constraint is no longer linear in all the unknown parameters.

Knowledge which is useful when fitting submodels concerns the type of join between the submodels. Hudson (1966) classifies the joins into three categories defined by the following cross-tabulation:

	$\hat{C}_j$ lies strictly between two successive values of X_i	$\hat{C}_j$ coincides with a value of X_i
Curves on either side of the join do not have equal slopes	Type One	Type Two
Curves join with equal slopes	Type Three	Type Two

Knowing the type of join and the successive values of data points X_i between which C_j lies facilitates the search for estimates of the parameters.

In the early 1960's the potential of PWLR models with unknown join points began to be recognized in model building. At that time both Robison (1964) and Hudson (1966) independently approached the problem from a statistical viewpoint. Prior to these studies Quandt (1958) used maximum likelihood (ML) techniques in estimating parameters of linear regressions obeying two separate regimes. However, Quandt's models did not incorporate the constraint that the overall function be continuous at the join point.

Robison, like Quandt, took an ML approach to the problem. Assuming that the error terms are normally distributed with constant unknown variances, the likelihood equation for the two-dimensional case with r segments has the form:

$$L(\underset{\sim}{\beta}, \sigma^2, \underset{\sim}{C}) = \prod_{k=1}^{r} \left\{ \prod_{i=N_{k-1}+1}^{N_k} (2\pi\sigma^2)^{-1} \exp \left[-(y_i - \beta_{k0} - \beta_{k1} x_i)^2 (2\sigma^2)^{-1} \right] \right\} \tag{1.10}$$

where N_k is the number of data points in the k-th segment and N_0 is defined to be 0.

When the N_k are known, the ML estimators of the join points are obtainable from the ML estimators of the $\underset{\sim}{\beta}_k$ by equating the sample regressions and solving for the points of intersection. When, in fact, the N_k are not known, but the total sample size $N = \sum_k N_k$ is known, the likelihood, $L(\underset{\sim}{\beta}, \sigma^2, \underset{\sim}{C})$, becomes a function of the N_k as well as $\underset{\sim}{\beta}$, σ^2

and $\underset{\sim}{C}$. In this case for every choice of N_k a corresponding set of $(\hat{\underset{\sim}{\beta}},\hat{\sigma}^2)$ must be calculated and the likelihood, L, computed. The set of N_k which produces the maximum L is then chosen and the join points are estimated from the corresponding parameter estimates as before. The number of computations for this iterative procedure is prohibitive as k, the number of segments, increases since all combinations of N_k must be taken into account. The number of possible combinations for which calculations are made can be narrowed down somewhat by taking only those combinations which are plausible. A further complication arises because Robison defines the domain of the likelihood function to be the entire $\underset{\sim}{\beta}$-space. This definition of the domain can result in complex-valued ML estimates of the C_k.

A different approach which Hudson (1966) uses to deal with fitting segmented regression curves when the join points must be estimated is that of ordinary least squares (OLS) estimation. It is a well known fact that OLS and ML estimators are equivalent under the assumption of normality of the error terms. The OLS estimators obtained by Hudson's technique are always ML estimators when normality is assumed. However, Hudson restricts the domain of the likelihood to that β-space for which $f_k(C_k,\beta_k) = f_{k+1}(C_k,\beta_{k+1})$ has at least one real root in C_k-space. This restriction of the domain avoids Robison's problem of complex-valued estimates of C_k.

For the OLS approach the overall residual sum of squares (RSS) for the model is:

$$RSS = R = \sum_{i=1}^{N} [y_i - f(x_i)]^2 \qquad (1.11)$$

where $f(x_i)$ is as defined in (1.1). In the two-dimensional case with two segments equation (1.11) becomes:

$$R = \sum_{i=1}^{N_1} [y_i - f_1(x_i;\underset{\sim}{\beta}_1)]^2 + \sum_{i=N_1+1}^{N} [y_i - f_2(x_i;\underset{\sim}{\beta}_2)]^2$$

$$= \sum_{1}^{N_1} [y_i - \beta_{10} - \beta_{11}x_i]^2 + \sum_{N_1+1}^{N} [y_i - \beta_{20} - \beta_{21}x_i]^2 \qquad (1.12)$$

R is to be minimized with respect to both the $\underset{\sim}{\beta}_k$ and C_k. The overall OLS estimates are denoted $\hat{\underset{\sim}{\beta}}_k$, $\hat{C}_k$ and $\hat{f}(\underset{\sim}{X})$. Local OLS estimators are denoted by $\underset{\sim}{\beta}^*_k$ and $\underset{\sim}{C}^*_k$ where $\underset{\sim}{\beta}^*_k$ is the value of β_k which minimizes:

$$\sum_{x_i \varepsilon I^*_k} [y_i - f_k(x_i;\underset{\sim}{\beta}_k)]^2 \qquad (1.13)$$

and I^*_k is a specified interval $I^*_k = [C^*_{k-1}, C^*_k]$. The minimum value of this sum (1.13) is denoted ρ^*_k. Notice that β^*_k and ρ^*_k vary with the choice of I^*_k.

For theoretical purposes Hudson defines four types of joins. In addition to the three types presented above a join of Type Four is defined by two submodels whose overall solution leads to $\hat{C}_k < x_1$ or $\hat{C}_k > x_n$. Clearly, the study of a complete model with joins of Type Four is equivalent to the study of a degenerate overall model with fewer submodels and no joins of Type Four. Thus, no further discussion is needed for this fourth type of join.

When a join is of Type I Hudson proves that the estimated $\hat{\underset{\sim}{\beta}}_k$ are equal to the corresponding local OLS estimates $\hat{\underset{\sim}{\beta}}^*_k$ where the required interval is $I^*_k = [\hat{C}_{k-1}, \hat{C}_k]$. Thus if a join is known to be of Type I and the interval where the join occurs is known then it is a simple matter to estimate the abscissa of the join point by minimizing the local residual sum of squares and solving for the intersection point of the regression lines. If the interval is unknown or the type of join is unknown then an iterative procedure is needed.

If a join is known to be of Type II, that is the abscissa of the join point corresponds to a data point rather than lying between two data points, then the overall solution for the $\hat{\underset{\sim}{\beta}}_k$ is written as modified local OLS estimates:

$$\hat{\underset{\sim}{\beta}}_k = \hat{\underset{\sim}{\beta}}_k^* + \Delta\hat{\underset{\sim}{\beta}}_k^* \tag{1.14}$$

where $\Delta\hat{\underset{\sim}{\beta}}_k^*$ is a correction term introduced to insure that the $\hat{\underset{\sim}{\beta}}_k$ satisfy the constraint:

$$f_k(x_i;\hat{\underset{\sim}{\beta}}_k) = f_{k+1}(x_i;\hat{\underset{\sim}{\beta}}_{k+1}) \tag{1.15}$$

This constraint is linear in the $\hat{\underset{\sim}{\beta}}_k$ so that the $\Delta\hat{\underset{\sim}{\beta}}_k$ are easily found using LaGrange multipliers. The actual calculations for $\Delta\hat{\underset{\sim}{\beta}}_k^*$ are given in Hudson's paper. Again if the type of join is not known or the data point where the join occurs is unknown then an iterative procedure must be used to obtain estimates for the parameters.

Join points of the third type do not occur if the segments are straight lines or constants since the overall curve is constrained to be smooth at the join. Hudson's discussion and treatment of this type of join is much briefer than for joins of Types I and II. Basically, he proposes that iterative procedures can be used to obtain a solution and that each case requires special treatment depending on the particular model being fit.

When the type of join is unknown an iterative search procedure must be employed to find the parameter estimates which minimize the RSS. If the segments are straight lines or constants then the joins of Type III are immediately eliminated and the search is limited to joins of Types I and II. For the sake of clarity, this discussion will be limited to a model with only k=2 segments. As was true with Robison's technique the number of computations becomes unwieldy as k increases.

Hudson recommends a two step procedure beginning with a search for Type I solutions. The first step involves choosing values for i and fitting the $\hat{\underset{\sim}{\beta}}{}_k^*(i)$ for each value of i. The estimate $\hat{C}_1$ is then found by solving:

$$f_1(C_i;\underset{\sim}{\beta}{}_1^*(i)) = f_2(C_i;\underset{\sim}{\beta}{}_2^*(i)) \tag{1.16}$$

for $C_1(i)$. If the curves join in the right place, i.e.

$$x_i < C_1(i) < x_{i+1} \tag{1.17}$$

then define:

$$T(i) = \rho_1^*(i) + \rho_2^*(i) \tag{1.18}$$

If the curves do not intersect or if they do not intersect in the right place then set $T(i) = \infty$. Repeat this procedure for all relevant values of i. Choose that critical value of i for which T(i) is minimum and take the corresponding $C_1(i)$ as $\hat{C}_1$, the estimate of the join point. If the join is of Type I then there will exist a $T(i) < \infty$ and no further search is needed. This step can also be used if the join is known to be Type I but the interval (x_i,x_{i+1}) which contains C_1 is unknown.

If, on the other hand, $T(i) = \infty$ for all i, then it is necessary to search for joins of Type II. In this second step, values of i are chosen and the necessary adjustments are made to the local OLS estimates from step one to ensure that the Type II constraint is met. The adjusted overall RSS is denoted by S(i). The critical value of i for which S(i) is minimum is then chosen and the corresponding data point, x_i, is taken to be the estimate of C_1. If the join is known to be of Type II and only i is unknown then Step 2 can be used to find the estimates which minimize at the residual sum of squares.

Step 3, looking for solutions of Type III, is necessary when it is known that the submodels are not constants or straight lines. In addition, it is only necessary to look for joins of Type III with $x_i < \hat{C}_1 < x_{i+1}$ if $T(i) = \infty$ and

$$R^*(i) = \rho_1^*(i) + \rho_2^*(i) < \min_j \; T(j), S(j) \quad . \tag{1.19}$$

Hudson and Robison present the most general studies of PWLR. Most of the work which has been done subsequently deals with specific cases or applications or with methods to improve upon the iterative procedures put forth by Hudson.

Bellman and Roth (1969) use dynamic programming to determine an optimal fit to data when join points are unknown. They give a numerical example to illustrate the technique, but no data applications are included. Hawkins (1976) gives two methods of fitting piecewise multiple regression models. One method is based on dynamic programming and the other is a hierarchical procedure. The approach of McGee and Carleton (1970) employs a combination of hierarchical clustering and standard regression theory to fit a piecewise continuous function. Unlike Hudson and Robison they do not require that the functions be continuous at the joins. Numerical applications include an analysis of the data created by Quandt (1958) in his discussion and also data from the financial world based on changes in the stock exchange. Gallant and Fuller (1973) look only at Hudson's Type III joins and propose a reparameterization which allows for the use of the Modified Gauss-Newton fitting technique. The technique is applied to data on preschool boys' weight-height ratio versus age. Both a linear-quadratic and a linear-quadratic-quadratic model are fit. Singpurwalla (1974) employs Hudson's technique for Type I or II situations which arise in accelerated life-testing experiments. He uses an iterative weighted least-squares (WLS) technique to take into account the heteroscedastic nature of the error terms. Ertel and Fowlkes (1976) develop an efficient algorithm for fitting PWLR models and also a plotting procedure which shows the existence and location of changes in the regression models. These authors compare their algorithm with the McGee-Carleton approach using generated samples. Hasselblad, Creason, and Nelson (1976) use a "hockey stick" function in a dose-response type situation when it is desired to estimate threshold levels of toxins. The term "hockey stick" is derived from the shape of the curve obtained when fitting a two-segment PWLR model which has a constant for the first segment and the constant plus a slope function as the second segment. The threshold level is estimated by the point where the regression changes from a constant -- i.e. the join point. In addition to an example, the PWLR model is compared to probit and logit models which are typically used in dose-response situations. Magoun (1978) extends the definition of join point to place emphasis on functional properties then employs an algorithm based on the numerical analysis technique of divided differences to estimate the join point. His definition of the join point is an abscissa point which is indicating a change in functional properties, that is, a critical point of inflection. His algorithm first tests for such an inflection point then uses the information gained from this test to estimate the PWLR model parameters. The algorithm is applied to the examples given by Hudson (1966) and in addition simulation studies were done to compare the results obtained from this procedure to those obtained using Hudson's iterative procedure. Lerman (1980) proposes a grid search method of fitting segmented curves. He compares this new procedure to the standard procedure proposed by Hudson and concludes that the grid search works for a wider range of models than the standard procedure.

1.3 METHODS OF DEALING WITH MEASUREMENT ERROR

1.3.1 Models

Piecewise linear regression, like the usual regression model, is based on the assumption that the independent variables are measured without error. It seems obvious that this assumption does not always hold in either regression situation. However, no work has been done to study the effects of measurement error when piecewise models are being used. The following review, therefore, deals with techniques currently employed with single function regression models. The model presented assumes only one independent variable, X, but it can easily be extended to more than one independent variable.

Suppose we are given variables X_i and Y_i related by the linear equation $Y_i = \beta_0 + \beta_1 X_i$ with both variables subject to random measurement error. Let $X_i^* = X_i + V_i$ and $Y_i^* = Y_i + U_i$ represent the actual observed values of these variables where V_i and U_i are the error terms and $E(V_i) =$

$E(U_i) = 0$. We have the usual regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where the disturbance term, ε_i, is usually thought of as representing explanatory variables which have not actually been included in the model. The ε_i are identically distributed with mean 0 and finite variance σ_ε^2 .

The disturbance term could also have a component representing measurement error in Y_i and the OLS estimates of the regression parameters would still be valid since the error term would still be independent of X_i. This model can be written as $Y_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i^*$ where $\varepsilon_i^* = \varepsilon_i + U_i$. The * notation for Y_i and ε_i will be dropped in the following discussion since it is not needed for clarity.

The measurement error problem becomes complicated with the additional knowledge that the X_i are measured with error. We then have the model:

$$Y_i = \beta_0 + \beta_1(X_i^* - V_i) + \varepsilon_i$$

which can be rewritten as:

$$Y_i = \beta_0 + \beta_1 X_i^* + \varepsilon_i - V_i\beta_1 \quad .$$

The error term, $\varepsilon_i - \beta_1 V_i$, is no longer independent of the imperfectly measured X_i^* since X_i^* is a function of V_i. It is a well known fact that applying OLS techniques in this situation leads to biased as well as inconsistent estimates of the regression parameters (Neter and Wasserman, 1974).

To summarize, we are looking at a model of the form:

$$Y_i = \beta_0 + \beta_1 X_i^* + \varepsilon_i - \beta_1 V_i \tag{1.20}$$

where the

(i) ε_i are identically distributed (i.d.) with mean 0 and finite variance σ_ε^2 which are uncorrelated and independent of X_i;

(ii) V_i are i.d. with mean 0 and finite variance σ_v^2 which are uncorrelated and independent of X_i;

(iii) ε_i and V_i are uncorrelated; and

(iv) true relation between X_i and Y_i is a linear one:

$$Y_i = \beta_0 + \beta_1 X_i \quad .$$

Classical approaches to the measurement error problem assume normally distributed error terms and require additional information in order to obtain consistent estimates of β_1 using ML techniques. Such additional information would be the values of σ_ε^2 or σ_v^2 or $\lambda = \sigma_\varepsilon^2/\sigma_v^2$. Madansky (1959) gives the estimates for these three cases as well as the overidentified case where both σ_ε^2 and σ_v^2 are known. He points out that the resulting estimates are not ML estimates and that their only known optimal property is consistency. He then raises the question concerning what other kinds of information besides the knowledge of σ_ε^2, σ_v^2 or λ could be used to obtain an estimate of β_1. He gives a comprehensive review of methods which require other kinds of information and presents an example comparing the various methods.

Cochran (1968) points out that in survey data, in particular, it may not be realistic to assume that the error terms V_i are uncorrelated with the true value X_i. In this situation, additional information about the covariance between V_i and X_i is needed in order to obtain estimates of the slope, β_1. Moran (1971) also discusses the classical approach by considering the normal distribution assumptions. In addition to cases where σ_ε^2, σ_v^2 and λ are known he looks briefly at the case where β_0 is known and then the cases where ε_i and V_i are correlated.

In the three sections that follow alternatives to the classical approach are given. These approaches do not require the distributional assumptions needed in the classical approach. Several grouping techniques will be discussed followed by a presentation of ways in which additional or "instrumental" variables can be used to eliminate biases due to measurement error. Finally estimation procedures which are used when observations are replicated are given.

1.3.2 Method of Grouping

Wald (1940) proposed a method of grouping which yields a consistent estimate of the slope even when the independent variable is measured with error. This method involves ordering the observations in increasing order of magnitude on the values of X_i^*, dividing the observations into two groups at $X_{(p_1)}$, say, where $p_1 = 1/2$ in this case, and treating the smaller observations as G_1 and the larger half of the observations as G_2. He shows that with the additional conditions:

(v) the groupings are independent of the error terms and

(vi) $\liminf\limits_{N\to\infty} \left(\sum\limits_{G_1} X_i - \sum\limits_{G_2} X_i\right)N^{-1} > 0$,

a consistent estimator of the slope can be calculated as:

$$b_1 = (\bar{Y}_1 - \bar{Y}_2)/(\bar{X}_1 - \bar{X}_2) \tag{1.21}$$

where $\bar{X}_1$ and $\bar{Y}_1$ denote the mean of the X and Y values for G_1, etc. Other variations on this grouping method have been proposed.

Bartlett (1949) determined that a more efficient estimator is obtained if the observations are divided into three equal groups ($p_1 = 1/3$)

rather than two groups as proposed by Wald. Bartlett's estimator makes use of the two extreme groups and discards the observations which fall in the middle group.

A third version of the grouping technique proposed by Nair and Banerjee (1943) divides the observations into quarters. The two extreme groups are again used to estimate the slope and the two middle groups of observations are discarded.

Another version, also based on four groups, was proposed by Lindley (1947). This method requires the calculation of two slope estimates. The first slope is estimated using the first and third quarter as the two groups and the second estimate makes use of the second and fourth groups. The mean of these two slope estimates is then taken as the final slope estimate.

In applying any of these grouping techniques the difficulty arises in ensuring that the assumptions are met. Madansky (1959) stresses that basing the ordering on the observed X*'s which are measured with error simulates ordering on the X's. However, it does not insure the consistency of the slope estimate since the method of grouping may not be independent of the errors. He claims that if there exists a δ such that $PR(|V| \geq \delta)$ is negligible and if there are a small number of X_i^*'s in $[X_{(p_1)} - \delta, X_{(p_1)} + \delta]$, then there is a high probability that the grouping based on the order of the X_i^*'s will be the same as the grouping based on the X_i's.

Blalock, Wells, and Carter (1970) point out that in the social sciences one virtually never finds data in which grouping on the measured X_i^* can be done in such a way that V_i is independent of the grouping since the measurement error is usually not small. Carter and Blalock (1970) perform simulation studies to compare the bias in the slope estimates with varied sizes of measurement error. They determine that the effect of this nonindependence of the error term is quite serious especially for the quality of data found in sociological studies. Their results showed Lindley's quarter grouping technique to be the least biased when the error terms were assumed to be normally distributed. However, the biases for all of the grouping techniques were on the same order as the bias for the OLS method.

Moran (1971) maintains that in order to use the methods correctly it is necessary to carry out the division of the observations into groups in such a way that the distributions of the error terms are unaffected. Under assumptions of normally distributed error terms he claims that grouping on the observed X_i^* values will alter the error distributions in such a way that Wald's conditions will not be met and thus a consistent slope estimate is unattainable. He concludes that situations in which all the conditions are satisfied must be rare.

1.3.3 Instrumental Variables

The use of instrumental variables (IV) was first developed by econometricians in solving the estimation problem when confronted with the error in variables model. Ideally, an IV is a third variable Z which (1) does not directly affect Y, but (2) is a reasonably direct cause of X, while (3) being uncorrelated with V and ε. Use of IV's, as found in current literature, falls into two categories. The first use of IV's involves a consideration of a simultaneous equations approach as originally proposed by the econometricians. The second of these methods relates back to the method of grouping and involves grouping observations on the exogenous variable Z_i.

Most of the early theory surrounding the use of IV's is to be found in economic literature. Johnston (1972) presents a general discussion of the theory of IV's then relates this theory to the situation where S is measured with error. Carleson, Sobel and Watson (1966) present an example from the biological sciences and urge biometricians to carefully consider the possibilities for IV's in biological research. Blalock, Wells, and Carter (1970) discuss the use of IV's from a sociological point of view.

The simultaneous equation idea behind the use of IV's is discussed by Durbin (1954), Johnston (1972), Madansky (1959), and Blalock, Wells, and Carter (1970). The equations involved are:

$$Y = \beta_0 + \beta_1 X + \epsilon_1 \qquad X = \alpha_0 + \alpha_1 Z + \epsilon_2 \tag{1.22}$$

The estimator

$$b_2 = \frac{\sum_i (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_i (Z_i - \bar{Z})(X_i^* - \bar{X}^*)} \tag{1.23}$$

is found by taking the ratio of the covariances of Z and Y and Z and X*. The estimator is consistent since the ratio of the covariances is not systematically affected by random measurement errors. Blalock, Wells, and Carter (1970) point out that it may be difficult, in practice, to simultaneously satisfy the three criteria for an ideal instrumental variable.

Durbin (1954) suggests ranking the X's in order and letting $Z_i = i$, the rank order of the observation. A further refinement of the use of rank orders is proposed for the case when the errors are thought to be so large that the rank ordering is affected by them. In this case, Durbin suggests arranging the X values according to magnitude into k groups and letting $Z_i = i$ for all the X's in the i-th group. A further possibility is to let Z_i equal some power of X. Madansky (1959) claims that using the rank order as the instrumental variable is more efficient than the Wald-Bartlett method of grouping. He also presents the formula for the approximate variance of b_2.

Durbin (1954) points out that by letting Z_i be equal to 1 for X greater than the median value, 0 for X values equal to the median and -1 for X values less than the median the Wald estimator is obtained. Bartlett's version of the grouping techniqe can be arrived at in a similar manner. However, defining the IV in this way does not eliminate the problems previously mentioned concerning the method of grouping.

Blalock, Wells, and Carter (1970) suggest that if an additional observed value Z_i is close to ideal, then ranking according to scores on Z_i and grouping by levels of Z should produce groups on X with similar scores. Thus by grouping on Z then computing b_1 using the means of the X's should take out the random error component. While the estimator is of the form seen in the grouping procedure it is based on the assumptions of the IV method -- namely, the _a priori_ assumption that Z can be used to group individuals who are similar with respect to X without distorting the relationship between Y and X which is measured by β_1. For this reason Blalock, Wells, and Carter (1970) claim that the resulting estimator should behave in the same way as the simultaneous equations approach estimator.

Moran (1971) stresses that the effectiveness of the IV approach depends on how strongly the X and Z are correlated. Thus it is still necessary to take into consideration how realistically the basic assumptions of the method are met.

1.3.4 Estimation with Replication of Observations

Suppose now that there are N_i observations X^*_{ij} on each of the n X_i's. The observed X values are now X^*_{ij} where $i = 1,2,\ldots,n$ and $j = 1,2,\ldots,N_i$. Given this kind of data Tukey (1951) and Housner and Brennan (1948) construct estimators for β_1. A good explication of these techniques is found in Madansky (1959).

The estimates proposed by Tukey (1951) can most simply be described by looking at the ANOVA table:

ANOVA Table

	Source	Mean Square	Expected Mean Square
Between	I	$\sum_{i=1}^{n} N_i(\bar{X}^*_i-\bar{X}^*)^2/(n-1)$	$\sigma_v^2 + [(N^2-\sum_i N_i^2)/(nN-N)]\sigma_x^2$
	II	$\sum_{i=1}^{n} N_i(\bar{X}^*_i-\bar{X}^*)(\bar{Y}_i-\bar{Y})/(n-1)$	$\sigma_{v\varepsilon} + [(N^2-\sum_i n_i^2)/(nN-N)]\beta\sigma_x^2$
	III	$\sum_{i=1}^{n} N_i(\bar{Y}-Y^{-})^2/(n-1)$	$\sigma_\varepsilon^2 + [(N^2-\sum_i N_i^2)/(nN-N)]\beta^2\sigma_x^2$
Within	IV	$\sum_i^N \sum_j^{N_i} (X^*_{ij}-\bar{X}^*_i)^2/(N-n)$	σ_v^2
	V	$\sum_i \sum_j (X^*_{ij}-\bar{X}^*_i)(Y_{ij}-\bar{Y}_i)/(N-n)$	$\sigma_{v\varepsilon}$
	VI	$\sum_i \sum_j (X_{ij}-\bar{Y}_i)^2/(N-n)$	σ_ε^2

where $\bar{X}^*_i = \sum_j X^*_{ij}/N_i$ $\quad \bar{Y}_i = \sum_j Y_{ij}/N_i$ $\quad N = \sum_i N_i$

$\bar{X}^* = \sum_i \sum_j X^*_{ij}/N$ $\quad \bar{Y} = \sum_i \sum_j Y_{ij}/N$

It can be seen that the ratios:

$$b_3 = (II - V)/(I - IV) \tag{1.24}$$

$$b_4 = (III - VI)/(II - V) \tag{1.25}$$

$$b_5 = \sqrt{\frac{III - VI}{I - IV}} \tag{1.26}$$

can be used to estimate β_1. When n and N_i tend to infinity then these estimators are obviously consistent.

Housner and Brennan (1948) propose another estimator for this type of data which Madansky (1959) claims can be considered as a variant of the method of grouping. Madansky's claim is based on studies of the optimal efficiency of this estimator. Housner and Brennan (1948) make use of the fact that a sample value of β_1 is:

$$b_{ijkl} = (Y_{ik} - Y_i)/(X^*_{ik} - X^*_j) \tag{1.27}$$

and that the error in variables model can be rewritten as:

$$(X^*_{ik} - X^*_j)b_{ijkl} = (X^*_{ik} - X^*_j)\beta_1 + (\varepsilon_{ik} - \varepsilon_j) - \beta_1(V_{ik} - V_j) \tag{1.28}$$

to arrive at the estimator:

$$b = \frac{\sum_{i=1}^{n} \bar{Y}_i N_i \left(\sum_{j=1}^{n} N_j - 2 \sum_{j=1}^{i} N_j + N_i \right)}{\sum_{i=1}^{n} \bar{X}^*_i N_i \left(\sum_{j=1}^{n} N_j - 2 \sum_{j=1}^{i} N_j + N_i \right)} \tag{1.29}$$

This estimator is consistent providing $N_i \to \infty$ for at least two distinct values of i.

Madansky (1959) claims that if it is truly realistic to assume that the relation between Y and X is linear then the Housner-Brennan estimator is preferred since consistency does not depend on n. On the other hand, if this assumption is questionable, and one is merely trying to approximate the function by a linear relation over a small range then it is advisable to have larger n at the expense of the N_i and use one of the Tukey estimators. Moran (1971) states that none of these estimators are optimal and suggests that one take the sums of squares to estimate σ_ε^2 and σ_v^2 then use these estimates in the classical estimate for the case when the two variances are known.

1.4 RESEARCH PROPOSAL

The main focus of the present work is to study the effects of measurement error on the join point estimate of a PWLR model. First, effects of measurement error are studied for the Hudson (1966) OLS estimation procedure. An alternative approach for estimating the join point in the presence of measurement error is proposed which applies the grouping slope estimation procedure of Wald (1940) to the piecewise problem.

In Chapter 2 the effect of measurement error on the OLS estimate is evaluated. The OLS slope parameters are known to be biased (Neter and Wasserman, 1974). A study is made of the bias in the OLS join point estimate through the use of Taylor series approximations to the expected value and variance. In addition, a simulation study is carried out on a numerical model in order to compare the empirical distribution with results found from the asymptotic theory. The asymptotic theory requires knowledge about the approximate location of the join point. Hence, a further consideration is the effect of misspecification of location on the resulting estimate.

In Chapter 3 it is proposed that the grouping-type estimator suggested by Wald (1940) be used to obtain consistent slope estimates which are then used in estimating the join point. A method is proposed for obtaining an approximation to the variance of the estimator. Since the data is not replicated it is impossible to estimate this variance directly. As an alternative, a numerical method proposed by Tepping (1968) which is based on a Taylor series expansion was applied to the join point estimator. In order to actually arrive at the variance estimate a pseudo replication technique which is based on the collapsed strata method (Cochran, 1977, p. 138) of sample survey methodology is applied to the numerical Taylor series. In addition, a simulation study is done to determine the empirical distribution of the join point estimate from Wald's technique.

Chapter 4 compares the results of estimating the join point by Hudson's technique and by Wald's technique. It is hoped that the Wald estimation approach will perform well in the presence of measurement error and provide a viable alternative to the OLS approach for estimating the join point of a PWLR model.

In Chapter 5 the case of join point estimation is briefly explored for replicated data. With replications it is possible to obtain estimates of the error variances. Three different consistent slope estimators have been proposed by Tukey (1951) for replicated data. All three of these estimators are based on the partitioned sums of squares. A simulation study is done using each of the three slope estimators to obtain a join point estimate.

In Chapter 6 the use of the Hudson and Wald techniques are illustrated with an application to data. The techniques are used in an attempt to model basal body temperature during a woman's menstrual cycle and to arrive at an estimate of the join point corresponding to the day on which a phase shift occurs.

The final chapter outlines topics for further research.

CHAPTER 2

EFFECTS OF MEASUREMENT ERRORS FOR JOIN POINT ESTIMATES IN PIECEWISE LINEAR REGRESSION MODELS USING HUDSON'S TECHNIQUE

2.1 INTRODUCTION

The usual assumption about the independent variables in a regression model as described in Chapter 1 underlies the development by Hudson (1966) of the join point estimate for a piecewise linear regression (PWLR) model. It is well known that when measurement error occurs in the independent variable the estimates of the regression coefficients obtained using ordinary least squares procedures are biased (Neter and Wasserman, p. 169). The question then arises about what effect this measurement error and the resulting regression coefficient bias have on the join point estimate when applying the analysis technique developed by Hudson. Since the join point is the ratio of differences between regression coefficients, the bias cannot be expected to be as simple as the bias of the regression coefficients themselves. The topic of this chapter is to study the behavior of the join point estimate from the Hudson technique when there is measurement error in the independent variable.

In order to study this behavior as the measurement error increases several approaches will be taken. First, approximate asymptotic properties of the estimator will be derived using Taylor series expansions. Next, the effects of increasing the sample sizes will be studied by comparing the asymptotic results for two samples. A simulation study will also be undertaken to obtain the empirical distribution of the join point estimate. For the sake of comparison the parameters estimated will be substituted into the asymptotic formulae. Finally, conclusions will be drawn based on the comparisons between the asymptotic results and the empirical results.

The specific model under study and a description of the estimation procedure is given in section 2.2. Section 2.3 presents the asymptotic properties of the estimator. The simulation procedure is discussed in section 2.4. Finally, the results from a numerical problem are given in section 2.5 and conclusions are drawn in section 2.6.

2.2 MODEL AND ESTIMATION PROCEDURE

For the purpose of this study one of the simplest cases of PWLR, two intersecting straight lines, will be used. The model itself is described in section 2.2.1 along with the necessary assumptions. The estimation procedure proposed by Hudson as it will be used here is described in section 2.2.2 .

2.2.1 Model Under Study

The model is of the form:

$$y_i = \begin{cases} \beta_{10} + \beta_{11}x_i - \varepsilon_i - \nu_i\beta_{11}\ , & X_i \leq \alpha \\ \\ \beta_{20} + \beta_{21}x_i - \varepsilon_i - \nu_i\beta_{2i}\ , & X_i \geq \alpha\ , \end{cases} \tag{2.1}$$

where $x_i = X_i + \nu_i$ is the observed value of the true X_i and ν_i is the measurement error; $y_i = Y_i + \varepsilon_i$ is the observed value of the true Y_i

and ε_i is the random error; and α is the abscissa of the join point such that $X_i \leq \alpha \leq X_{i+1}$. In addition, there is the important constraint that the two segments intersect at α. This constraint can be expressed as:

$$\beta_{10} + \beta_{11}\alpha = \beta_{20} + \beta_{21}\alpha \tag{2.2}$$

For theoretical developments the following assumptions will be made:

(i) $i = N_1$ such that $X_i \leq \alpha < X_{i+1}$ is known.

(ii) X_i are fixed with mean $\bar{X}$ and variance s_X^2.

(iii) ε_i are iid with mean 0 and variance σ_ε^2.

(iv) ν_i are iid with mean 0 and variance σ_ν^2.

(v) X_i, ε_i, and ν_i are uncorrelated.

(vi) $Y_i = \begin{cases} \beta_{10} + \beta_{11}X_i\ , & X_i \leq \alpha\ , \\ \beta_{20} + \beta_{21}X_i\ , & X_i \geq \alpha \end{cases}$

is the true relation between X and Y.

Note that when $\sigma_\nu^2 = 0$ the model is the usual PWLR model with fixed X_i.

When measurement error is present, the error term in the model has an additional component, $\nu_i\beta_{k1}$, $k = 1, 2$. With this component different from 0, the error term is no longer independent of the regression parameters.

2.2.2 Hudson's Estimation Procedure

An estimation procedure which is based on the assumption that there is no measurement error is described in Chapter 1. The highlights of that procedure are reiterated here as the concept of measurement error is introduced into the model.

According to Hudson (1966), the join point in a straight line model with no measurement error can be of two types. If α, the abscissa of the join point, falls between two X-values, then the join is said to be a Type I join. If α corresponds to one of the X-values, i.e. $\alpha = X_i$, for some $2 \leq i \leq N-1$, then the join is a Type II join. When the join is of the second type, a slight modification must be made in the OLS estimators of the regression parameters in order to satisfy the additional constraint that $\alpha = X_i$. The procedure when N_1 is known is described in section 2.2.2a for both the Type I and Type II joins. For the case where N_1 is not known, the iterative procedure is described briefly in section 2.2.2b. For a more detained description of the procedures see chapter 1.

2.2.2a <u>Procedure when N_1 Is Known</u>. Assume first that the join is of Type I and that the interval where the join occurs, (X_i, X_{i+1}) is known, although the actual point α is unknown. Two submodels are determined in the Hudson procedure by minimizing the "local" residual sums of squares (RRS). The first local segment is determined by $(x_1, x_2, \ldots, x_i)$ and the second local segment by $(x_{i+1}, \ldots, x_N)$. The join point is then estimated as the point of intersection of the two regressions, namely:

$$\hat{\alpha} = (\hat{\beta}_{10} - \hat{\beta}_{20})/(\hat{\beta}_{21} - \hat{\beta}_{11}) \ . \qquad (2.3)$$

Assume now that the join is of the second type and $i = N_1$ is known. The problem then reduces to a model linear in the β_k and the solution is trivial.

2.2.2b <u>Procedure when N_1 is Unknown</u>. If $i = N_1$ is unknown, then the following iterative procedure must be used. For each relevant value of i, the local RSS's are minimized and the overall RSS is calculated as the sum of the two local RSS's. The value of i for which $\hat{\alpha}(i)$ falls between x_i and x_{i+1} and for which the overall RSS is minimum is determined. The corresponding submodels are then used to estimate $\hat{\alpha}$.

When $i = N_1$ is unknown, the same iterative procedure is used as for a Type I join except that the estimates of the β_k and the local RSS are modified to meet the extra constraint that the join corresponds to a data point.

If neither the type of join nor $i = N_1$ is known (the most realistic case), then the iterative search begins by looking for joins of Type I. For each relevant value of i, the corresponding estimate $\hat{\alpha}(i)$ of α is calculated. If $x_i < \hat{\alpha}(i) < x_{i+1}$, then the overall RSS is recorded as T(i). If $\hat{\alpha}(i) \notin (x_i, x_{i+1})$, then set $T(i) = \infty$. If the join is a Type I join, then there will be a $T(i) < \infty$ and the min T(i) will correspond to the desired solution. If $T(i) = \infty$ for all i, the search must proceed for a Type II join by modifying the estimated regression parameters and determining the i which leads to the minimum modified RSS. By setting $\hat{\alpha}$ equal to the value of the i^{th} data point, x_i gives an estimate of α.

2.3 ASYMPTOTIC PROPERTIES OF HUDSON ESTIMATE UNDER THE MEASUREMENT ERROR MODEL

In studying an estimation procedure it is desirable to know the distribution of the estimator or at least how the first two moments of the estimator behave. In this section the expected values and variances are explored with the help of Taylor series approximations. In addition a study is made of the expected value as a function of the measurement error variance.

The estimator $\hat{\alpha}$ given by expression 2.3 of the join point of the two-segment PWLR model is a non-linear function of the regression parameters. Since it is difficult to determine the distribution of $\hat{\alpha}$, first order Taylor series expansions are used in order to obtain a linear approximation to $E(\hat{\alpha})$. Taylor series expansions are also used in deriving an approximation to the variance or mean squared error (MSE) of the estimator $\hat{\alpha}$. If $\hat{\alpha}$ is an unbiased or consistent estimator of α, as in the case of no measurement error, then the Taylor series expansion approximates the variance. However, in the case where $\hat{\alpha}$ is a biased estimate, the MSE is obtained rather than the variance, where the MSE is the variance plus the square of the bias. This distributional theory is conditioned on the correct choice of i. Hence, the assumption that i is known is implicit in the derivations which follow. Section 2.3.1 gives the Taylor series approximation to the expected value and mean squared error in addition to the derivations of the terms which make up these approximations. The last section (2.3.2) discusses the behavior of the expected value as a function of the measurement error variance.

2.3.1 Derivation of Expected Value and Mean Squared Error

The Taylor series expansion of the function $f(\hat{\underset{\sim}{\beta}}) = \hat{\alpha}$ leads to the following approximations to the expected value and MSE of $\hat{\alpha}$ (Kendall and Stuart, 1977, p. 246):

$$E(\hat{\alpha}) = \frac{E(\text{Num})}{E(\text{Den})} + O(n^{-1}) \qquad (2.4)$$

such that for this ratio estimator term

$$O(n^{-1}) = \frac{E(\text{Num})}{E(\text{Den})}\left\{\frac{\text{Var}(\text{Den})}{[E(\text{Den})]^2} - \frac{\text{Cov}(\text{Num},\text{Den})}{E(\text{Num})E(\text{Den})}\right\} \qquad (2.4a)$$

where $\text{Num} = \hat{\beta}_{10} - \hat{\beta}_{20}$ and $\text{Den} = \hat{\beta}_{21} - \hat{\beta}_{11}$, and n is the sample size. Similarly,

$$\mathrm{MSE}(\hat{\alpha}) = \left[\frac{E(\mathrm{Num})}{E(\mathrm{Den})}\right]^2 \left\{\frac{\mathrm{Var(Num)}}{[E(\mathrm{Num})]^2} + \frac{\mathrm{Var(Den)}}{[E(\mathrm{Den})]^2} - \frac{2\mathrm{Cov(Num,Den)}}{E(\mathrm{Num})E(\mathrm{Den})}\right\} + o(n^{-1}) \tag{2.5}$$

where $o(n^{-1})$ goes to 0 as n goes to ∞.

It is necessary to introduce additional notation at this point in order to clarify the derivations which follow. Without loss of generality, the X_i are ordered $X_1 < X_2 < \cdots < X_N$. Let $N = N_1 + N_2$ be the total sample size, where N_1 is the number of observations in the first segment -- those observations with abscissae less than or equal to α -- and N_2 is the number of observations in the second segment. Thus there are two segments with the following statistics defined for each $k = 1,2$ segment. Let

$$\bar{X}_k = \sum^{N_k} X_i/N_k \quad \text{and} \quad s^2_{X(k)} = \sum^{N_k} (X_i - \bar{X}_k)^2/N_k$$

be the mean and variance of the true X values. Also let

$$\bar{x}_k = \sum^{N_k} x_i/N_k \quad \text{and} \quad s^2_{x(k)} = \sum^{N_k} (x_i - \bar{x}_i)^2/N_k$$

be the mean and variance of the observed x values. The statistics $\bar{Y}_k$, $s^2_{Y(1)}$, $\bar{y}_k$, $s^2_{y(1)}$ are similarly defined. In addition, define

$$z_i = \frac{(x_i - \bar{x}_k)}{\sum^{N_k} (x_i - \bar{x}_k)^2}, \quad 1 \leq i \leq N,\ k = 1,2. \tag{2.6}$$

Thus, the OLS estimates of the regression parameters are:

$$\hat{\beta}_{k1} = \sum^{N_k} z_i y_i \tag{2.7}$$

and

$$\hat{\beta}_{k0} = \bar{y}_k - \hat{\beta}_{k1}\bar{x}_k , \quad k = 1,2. \tag{2.8}$$

The following lemmas are needed to calculate the approximate expectation and mean squared error of $\hat{\alpha}$:

<u>Lemma 2.3.1(a)</u> For large N_k,

$$E(\hat{\beta}_{k1}) \simeq \beta_{k1} \left\{\frac{s^2_{X(k)}}{s^2_{X(k)} + \sigma^2_\nu}\right\} , \quad k = 1,2. \tag{2.9}$$

<u>Proof</u>: For $k = 1,2$, conditioning on x, the left hand side of (2.9) can be written as

$$E\,\{E(\hat{\beta}_{k1}|x)\} = E\,\{E(\sum^{N_k} z_i y_i|x)\} . \tag{2.10}$$

Since z_i is constant given x, equation (2.10) becomes

$$E\,\{\sum^{N_k} z_i E(y_i|x)\} = E\,\{\sum^{N_k} z_i(\beta_{k0} + \beta_{k1}x_i - \nu_i\beta_{k1})\} . \tag{2.11}$$

It can be shown that $\Sigma z_i = 0$ and $\Sigma x_i z_i = 1$.

Hence, substituting in equation (2.11) and simplifying leads to:

$$E\,\{\beta_{k1} - \beta_{k1}\Sigma\nu_i z_i\}$$

$$= \beta_{k1} - \beta_{k1}\, E \left\{\frac{\Sigma\nu_i(x_i-\bar{x}_k)}{\Sigma(x_i-\bar{x}_k)^2}\right\} . \tag{2.12}$$

Assuming large N_k, the right hand side of equation (2.12) is approximately equal to:

$$\beta_{kl} - \beta_{kl} E \left\{ \frac{\sigma_{XV} + \sigma_{\nu}^{2}}{s_{X(k)}^{2} + \sigma_{\nu}^{2}} \right\} \quad , \tag{2.13}$$

where σ_{XV} represents cov(X,V).

But X and V are uncorrelated and all the terms under the expectation are constants; with sóme algebraic manipulations it can be shown that expression (2.13) simplifies to:

$$\beta_{kl} \left\{ \frac{s_{X(k)}^{2}}{s_{X(k)}^{2} + \sigma_{\nu}^{2}} \right\} \quad , \; k = 1,2.$$

q.e.d.

<u>Lemma 2.3.1(b)</u> For large N_k,

$$V(\hat{\beta}_{kl}) \approx \frac{\beta_{kl}\sigma_{\nu}^{2} + \sigma_{\varepsilon}^{2}}{N_k(s_{X(k)}^{2} + \sigma_{\nu}^{2})} \tag{2.14}$$

<u>Proof</u>: Conditioning on x, the left hand side of equation (2.14) can be written as:

$$V\{E(\hat{\beta}_{kl}|x)\} + E\{V(\hat{\beta}_{kl}|x)\} \quad . \tag{2.15}$$

From the left hand side of equation (2.12) it can be seen that

$$E(\hat{\beta}_{kl}|x) = \beta_{kl} - \beta_{kl}\Sigma\nu_i z_i \tag{2.16}$$

Now, $\nu_i z_i$ and $\nu_j z_j$ are uncorrelated when $i \neq j$ and also ν_i and z_i are uncorrelated since ν and X are uncorrelated; hence the variance of equation (2.16) can be written as:

$$\hat{\beta}^2_{k1}\ \Sigma\{V(\nu_i)V(z_i) + V(\nu_i)[E(z_i)]^2 + V(z_i)[E(\nu_i)]^2\}\ . \qquad (2.17)$$

But $V(\nu_i) = \sigma^2_\nu$ and $E(\nu_i) = 0$. So substituting into expression (2.17) and simplifying leads to:

$$\beta^2_{k1}\ \Sigma\sigma^2_\nu\ E(z^2_i)\ . \qquad (2.18)$$

After interchanging Σ and E it can be shown that

$$\Sigma(z^2_i) = \{1/\Sigma(x_i - \bar{x}_k)^2\}\ . \qquad (2.19)$$

Hence it follows from expressions (2.18) and (2.19) that the first term of expression (2.15) is

$$\beta^2_{k1}\sigma^2_\nu\ E\ \{1/\Sigma(x_i - \bar{x}_k)^2\}\ . \qquad (2.20)$$

Now it is well known that when x is fixed

$$V(\hat{\beta}_{k1}|x) = \sigma^2_\varepsilon/\Sigma(x_i - \bar{x}_k)^2\ . \qquad (2.21)$$

Substituting expressions (2.20) and (2.21) into expression (2.15) and simplifying produces

$$[\beta^2_{k1}\sigma^2_\nu + \sigma^2_\varepsilon]\ E\ \{1/\Sigma(x_i - \bar{x}_k)^2\}\ . \qquad (2.22)$$

Assuming large N_k, expression (2.22) is approximately equal to

$$(\beta^2_{k1}\sigma^2_\nu + \sigma^2_\varepsilon)\ E\ \{1/[N_k(s^2_{X(k)} + \sigma^2_\nu)]\}\ . \qquad (2.23)$$

Since all the terms under the expectations are constants, expression (2.23) is equal to the right hand side of equation (2.14).

q.e.d.

Corollary 1:

If $\sigma_\nu^2 = 0.0$ then $E(\hat{\beta}_{k1}) = \beta_{k1}$ (2.24)

$$\text{and} \quad V(\hat{\beta}_{k1}) = \frac{\sigma_\varepsilon^2}{N_k s_{X(k)}^2} \tag{2.25}$$

Proof: By substitution.

Lemma 2.3.2(a)

$$E(\hat{\beta}_{k0}) = \beta_{k0} + \frac{\beta_{k1}\bar{X}_k\sigma_\nu^2}{s_{X(k)}^2+\sigma_\nu^2} \tag{2.26}$$

Proof: Substituting equation (2.8) for $\hat{\beta}_{k0}$ and conditioning on x gives

$$E(\hat{\beta}_{k0}) = E[E(\bar{y}_k - \hat{\beta}_{k1}\bar{x}_k|x)] \quad . \tag{2.27}$$

It can be shown that the conditional expectation on the right hand side of (2.27) simplifies to

$$\beta_{k0} + \beta_{k1}\bar{x}_k\Sigma\nu_i z_i \quad . \tag{2.28}$$

Assuming large values of N_k, expression (2.28) is approximately equal to

$$\beta_{k0} + \frac{\beta_{k1}\bar{X}_k\sigma_\nu^2}{s_{X(k)}^2+\sigma_\nu^2} \quad . \tag{2.29}$$

Since all the terms in expression (2.29) are constants, taking the expected value over x leads to the same equation.

q.e.d.

Lemma 2.3.2(b)

$$V(\hat{\beta}_{k0}) \approx \beta_{k1}^2 \left\{ \frac{\sigma_\nu^2}{N_k(s_{X(k)}^2+\sigma_\nu^2)} \left[\frac{\sigma_\nu^2}{N_k} + \frac{(\sigma_\nu^2)^2}{s_{X(k)}^2+\sigma_\nu^2} + \bar{\bar{X}}_k^2 \right] \right\}$$

$$+ \frac{\sigma_\varepsilon^2}{N_k}\left[1 + \frac{\bar{\bar{X}}_k^2}{s_{X(k)}^2+\sigma_\nu^2}\right] \qquad (2.30)$$

Proof: Conditioning on x, $V(\hat{\beta}_{k0})$ is equal to

$$V[E(\hat{\beta}_{k0}|x)] + E[V(\hat{\beta}_{k0}|x)] \quad . \qquad (2.31)$$

From expression (2.28) it can be seen that

$$E(\hat{\beta}_{k0}|x) = \beta_{k0} + \beta_{k1}\bar{x}_k\Sigma\nu_i z_i \quad . \qquad (2.32)$$

Since β_{k0} and β_{k1} are constant, the variance of equation (2.32) can be written as

$$\beta_{k1}^2 V(\bar{x}_k\Sigma\nu_i z_i) \quad , \qquad (2.33)$$

but expression (2.33) is equal to

$$\beta_{k1}^2\{V(\bar{x}_k)V(\Sigma\nu_i z_i) + V(\bar{x}_k)[E(\Sigma\nu_i z_i)]^2$$

$$+ V(\Sigma\nu_i z_i)[E(\bar{x}_k)]^2\} \qquad (2.34)$$

It can be seen from the proof of Lemma 2.3.1(b) that

$$V(\Sigma\nu_i z_i) \approx \frac{\sigma_\nu^2}{N_k(s_{X(k)}^2+\sigma_\nu^2)} \qquad (2.35)$$

when N_k is large. It can also be easily shown that for large N_k ,

$$E(\Sigma \nu_i z_i) \simeq \frac{\sigma_\nu^2}{s_{X(k)}^2+\sigma_\nu^2} \tag{2.36}$$

and

$$V(\bar{x}_k) = \sigma_\nu^2/N_k \tag{2.37}$$

$$E(\bar{x}_k) = \bar{X}_k \tag{2.38}$$

Now, substituting equations (2.35), (2.36), (2.37), and (2.38) into expression (2.34) and simplifying give

$$\beta_{k1}^2 \left\{ \frac{\sigma_\nu^2}{N_k(s_{X(k)}^2+\sigma_\nu^2)} \left[\frac{\sigma_\nu^2}{N_k} + \frac{(\sigma_\nu^2)^2}{s_{X(k)}^2+\sigma_\nu^2} + \bar{X}_k^2 \right] \right\} . \tag{2.39}$$

Now, when x is fixed, it is well known that

$$V(\hat{\beta}_{k0}|x) = \sigma_\varepsilon^2 \left[\frac{1}{N_k} + \frac{\bar{x}_k^2}{\Sigma(x_i-\bar{x}_k)^2} \right] \tag{2.40}$$

but when N_k is large, equation (2.40) is approximately equal to

$$\frac{\sigma_\varepsilon^2}{N_k} \left[1 + \frac{\bar{X}_k^2}{s_{X(k)}^2+\sigma_\nu^2} \right] . \tag{2.41}$$

Hence, substituting expressions (2.39) and (2.41) into expression (2.31) gives the desired result.

q.e.d.

Corollary 2: If $\sigma_\nu^2 = 0.0$

$$E(\beta_{k0}) = \beta_{k0} \tag{2.42}$$

$$\text{and} \quad V(\hat{\beta}_{k0}) = \frac{\sigma_\varepsilon^2}{N_k} \left[1 + \frac{\bar{X}_k^2}{s_{X(k)}^2} \right] . \tag{2.43}$$

Proof: By substitution.

<u>Lemma 2.3.3</u>

$$\mathrm{Cov}(\hat{\beta}_{k0},\hat{\beta}_{k1}) = \frac{-\sigma_\varepsilon^2\bar{X}_k}{N_k(s^2_{X(k)}+\sigma^2_\nu}\qquad(2.44)$$

<u>Proof</u>: By definition the $\mathrm{cov}(\hat{\beta}_{k0},\hat{\beta}_{k1})$ can be written as

$$E[E(\hat{\beta}_{k0}\hat{\beta}_{k1}|x)] - E(\hat{\beta}_{k0})E(\hat{\beta}_{k1}) \quad .\qquad(2.45)$$

Substituting equations (2.9) and (2.26) into the second term of expression (2.45) gives

$$\frac{\beta_{k1}s^2_{X(k)}}{s^2_{X(k)}+\sigma^2_\nu}\left(\beta_{k0}+\frac{\beta_{k1}\bar{x}_k\sigma^2_\nu}{s^2_{X(k)}+\sigma^2_\nu}\right) \quad .\qquad(2.45a)$$

Using the definition of $\hat{\beta}_{k0}$, equation (2.8), in the first term of expression (2.45) leads to

$$E\{E[(\bar{y}_k - \hat{\beta}_{k1}\bar{x}_k)\hat{\beta}_{k1}|x]\} \quad .\qquad(2.46)$$

After multiplying through by $\hat{\beta}_{k1}$, expression (2.46) becomes

$$E[E(\hat{\beta}_{k1}\bar{y}_k|x) - \bar{x}_kE(\hat{\beta}^2_{k1}|x)] \quad .\qquad(2.47)$$

But $\hat{\beta}_{k1}$ and $\bar{y}_k$ are uncorrelated (cf. Draper and Smith, p. 22); hence, the first conditional expectation of expression (2.47) becomes

$$E(\hat{\beta}_{k1}|x)E(\bar{y}_k|x)\qquad(2.48)$$

where $E(\hat{\beta}_{k1}|x) = \beta_{k1} - \beta_{k1}\Sigma\nu_iz_i$ (2.49)

and $E(\bar{y}_k|x) = \beta_{k0} + \beta_{k1}\bar{x}_k$. (2.50)

Now $E(\hat{\beta}^2_{k1}|x) = V(\hat{\beta}_{k1}|x) + [E(\hat{\beta}_{k1}|x)]^2$. (2.51)

Hence, substituting equations (2.49) and (2.50) into expression (2.48), then into expression (2.47), and substituting equations (2.49) and (2.40) into equation (2.51), then into expression (2.47) and simplifying, gives

$$E\{\beta_{k0}\beta_{k1} - \beta_{k0}\beta_{k1}\Sigma\nu_i z_i + \beta^2_{k1}\bar{x}_k\Sigma\nu_i z_i - \frac{\sigma^2_\varepsilon\bar{x}_k}{\Sigma(x_i-\bar{x}_k)^2} - \beta^2_{k1}\bar{x}_k(\Sigma\nu_i z_i)^2\} \quad . \tag{2.52}$$

Assuming N_k is large, expression (2.52) is approximately equal to

$$E\left\{\beta_{k0}\beta_{k1} - \frac{\beta_{k0}\beta_{k1}\sigma^2_\nu}{s^2_{X(k)}+\sigma^2_\nu} + \frac{\beta^2_{k1}\bar{X}_k\sigma^2_\nu}{s^2_{X(k)}+\sigma^2_\nu} - \frac{\sigma^2_\varepsilon\bar{X}_k}{N_k(s^2_{X(k)}+\sigma^2_\nu)} - \beta^2_{k1}\bar{X}_k\left(\frac{\sigma^2_\nu}{s^2_{X(k)}+\sigma^2_\nu}\right)^2\right\} \tag{2.53}$$

Finally, substituting expressions (2.53) and (2.45a) into expression (2.45), taking expectations, and simplifying, leads to the desired result.

q.e.d.

Using the above lemmas and the fact that for $k = 1,2$,

$$E(\hat{\beta}_{ik} - \hat{\beta}_{jk}) = E(\hat{\beta}_{ik}) - E(\hat{\beta}_{jk}) \tag{2.54}$$

$$V(\hat{\beta}_{ik} - \hat{\beta}_{jk}) = V(\hat{\beta}_{ik}) + V(\hat{\beta}_{jk}) \tag{2.55}$$

and

$$Cov(\hat{\beta}_{10}-\hat{\beta}_{20}, \hat{\beta}_{21}-\hat{\beta}_{11}) = -Cov(\hat{\beta}_{10},\hat{\beta}_{11}) - Cov(\hat{\beta}_{20},\hat{\beta}_{21}) \tag{2.56}$$

the approximate expected value and mean squared error can be expressed as functions of the model parameters.

2.3.2 Expected Value as a Function of Measurement Error Variance

From equation (2.4), the asymptotic expected value of $\hat{\alpha}$ can be expressed as

$$E(\hat{\alpha}) \approx \frac{E(\hat{\beta}_{10} - \hat{\beta}_{20})}{E(\hat{\beta}_{21} - \hat{\beta}_{11})} \tag{2.57}$$

since the right hand side of (2.4a) goes to 0 as $N_1, N_2 \to \infty$.

By applying lemmas 2.3.1(a) and 2.3.2(a), equation (2.57) can be rewritten as

$$E(\hat{\alpha}) \approx \tag{2.58}$$

$$\frac{(\beta_{10}-\beta_{20})(s^2_{X(1)}+\sigma^2_\nu)(s^2_{x(2)}+\sigma^2_\nu) + \beta_{11}\bar{X}_1\sigma^2_\nu(s^2_{X(2)}+\sigma^2_\nu) - \beta_{21}\bar{X}_2\sigma^2_\nu(s^2_{X(1)}+\sigma^2_\nu)}{\beta_{21}s^2_{X(2)}(s^2_{X(1)}+\sigma^2_\nu) - \beta_{11}s^2_{X(1)}(s^2_{X(2)}+\sigma^2_\nu)}$$

By looking at $E(\hat{\alpha}) = f(\sigma^2_\nu)$ as a function of σ^2_ν and taking derivatives of $f(\sigma^2_\nu)$ with respect to σ^2_ν the behavior of $\hat{\alpha}$ can be studied as σ^2_ν increases. The join point estimate is now of the form

$$f(V) = \frac{AV^2 + BV + C}{DV + E} \tag{2.59}$$

where

$$V = \sigma^2_\nu$$

$$A = (\beta_{10}-\beta_{20}) + \beta_{11}\bar{X}_1 - \beta_{21}\bar{X}_2 \tag{2.60}$$

$$B = (\beta_{10}-\beta_{20})(S_1+S_2) + \bar{X}_1S_2\beta_{11} - \bar{X}_2S_1\beta_{21} \tag{2.61}$$

$$C = (\beta_{10}-\beta_{20})S_1S_2 \tag{2.62}$$

$$D = \beta_{21}S_2 - \beta_{11}S_1 \tag{2.63}$$

$$E = S_1S_2(\beta_{21}-\beta_{11}) \tag{2.64}$$

$$S_1 = s^2_{X(1)}$$
$$S_2 = s^2_{X(2)}$$

The function f(V) is equal to the true value of α when $f(V) = C/E$ or equivalently when

$$V = 0 \quad \text{or} \qquad V = \frac{-(BE-CD)}{AE} \tag{2.65}$$

In order to determine the behavior of $\hat{\alpha}$ as σ^2_ν increases it is necessary to find the critical point(s) of the function f(V) by setting the first derivative equal to zero and solving for $V = V_o > 0$. To determine whether the critical point is a max or min involves study of the second derivative at the point V_o such that $f'(V_o) = 0$.

The first and second derivatives of f(V) with respect to V can then be written

$$f'(V) = \frac{ADV^2 + 2AEV + BE - CD}{(DV + E)^2} \tag{2.66}$$

$$f''(V) = \frac{2AE^2 - 2BED + 2CD^2}{(DV + E)^3} \tag{2.67}$$

<u>Lemma 2.3.4</u>

$$V_o = \frac{-E}{D} \pm \sqrt{\frac{E^2}{D} - \frac{BE - CD}{AD}} \tag{2.68}$$

<u>Proof</u>: In order for equation (2.66) to equal 0 the numerator must equal 0. Hence setting the numerator equal to 0 gives

$$ADV^2 + 2AEV + (BE - CD) = 0 \; .$$

Applying the formula for the solution of a quadratic and taking the positive solution, if it exists, gives the desired result. It is necessary to choose the positive solution since V is the measurement error variance which cannot be negative.

Lemma 2.3.5

$$f''(V_o) = \frac{2A}{[E^2 - (D/A)(BE - CD)]^{1/2}} \tag{2.69}$$

Proof: Substituting (2.68) into (2.67) and simplifying gives

$$f''(V_o) = \frac{2A}{[E^2 - (D/A)(BE - CD)]^{1/2}}$$

q.e.d.

Both V_o and $f''(V_o)$ are complicated functions of $\underset{\sim}{\beta}$, $\bar{X}_1$, $\bar{X}_2$, $s^2_{X(1)}$, and $s^2_{X(2)}$ and, therefore, dependent on the particular model being studied and the distribution of the X-values.

2.4 SIMULATION PROCEDURE FOR THE EMPIRICAL DISTRIBUTION OF THE JOIN POINT ESTIMATE

The purpose of this experiment was to study the empirical distribution of $\hat{\alpha}$ using Hudson's technique to estimate the join point when measurement error is present in the independent variable. Ideally, one would like to obtain a closed form for the distribution of the join point estimate. Since it was not possible to find such a closed form, approximations to the first two moments were determined in section 2.3. In this section a simulation procedure is set forth which produces the empirical distributions of the parameter estimates by looking at the estimates from a large number of samples. Theoretically, the mean and variance of an estimate over all the simulated samples should be close to the approximate moments obtained using asymptotic theory (section 2.3).

Section 2.4.1 gives a description of the procedure used to simulate the model given in section 2.1.1. Next, section 2.4.2 discusses the simulation of the assumption that N_1 is known and section 2.4.3 describes how the simulation is changed for the case when N_1 is unknown. Section 2.4.4 describes how the empirical moments were obtained from the simulated samples. Finally, section 2.4.5 describes the use of the simulation to obtain approximations of the asymptotic expected value and mean squared error.

2.4.1 Simulation of the Model

First, a mathematical model was specified corresponding to assumption (vi) of the statistical model. Thus the true values, the true regression parameters and the true join point were specified. The abscissa values were preselected to be one unit apart. The simulation procedure then generated N random error terms ϵ_i. These errors were generated as random normal deviates using the FORTRAN IV subroutine GAUSS which requires the mean and standard deviation to be specified. The ϵ_i were then added to the Y_i calculated from the mathematical model to produce y_i and a statistical model with random error but no measurement error. A second set of normal deviates corresponding to ν_i, the measurement error, was then generated in the same way. These random normal deviates had mean 0 and variance σ_ν^2. The ν_i values were then added to the X_i values to produce x_i, the so-called observed values. The x_i and y_i values were sorted on x_i and fed into a subroutine based on Hudson's algorithm to obtain estimates of $\hat{\alpha}$.

2.4.1a Simulation of N_1 Known. In order for the simulation results to be comparable to the asymptotic results it is necessary to have some prior knowledge of the approximate location of the join point. This knowledge is needed because one of the assumptions underlying the asymptotic theory is that N_1 is known -- where N_1 is such that $X_{N_1} \leq \alpha < X_{N_1+1}$. (Note that it is not the value of α itself which is known but between two which values α lies.) When there is no measurement error and thus the ordered X values are fixed knowing N_1 is equivalent to knowing X_{N_1} and there is no problem simulating the results. However, in this case, when measurement error is present, the concept of N_1 known is not so simple. While the X_i are fixed, x_i has a random component which may cause x_{N_1}, based on the ordered x_i, to correspond to a different value of X rather than X_{N_1}.

The approach taken to solve this problem in the simulation study was to assume that the prior knowledge is the value of x_{N_1}, which defines the interval $[x_{N_1}, x_{N_1+1}]$ containing α. Thus the observed x_{N_1} are taken to be in the first segment while those values of x_i greater than x_{N_1} are taken to be in the second segment. An additional dimension is thus added to the study by observing the effects of misspecification of x_{N_1} on the parameter estimates.

As a first step in the study of the Hudson procedure in the face of measurement error, the cut point was chosen as the true join point value in order to minimize the number of observed data points which are classified in the wrong segment in any one sample. Obviously, if the true value is known there are much simpler techniques for fitting a segmented model (Neter and Wasserman, 1974; Monti, Koch, and Sawyer, 1978).

2.4.1b Simulation of N_1 Unknown. When there is no prior knowledge of the location of the join point -- a more realistic situation -- the simulated (x_i, y_i) observations are fed into a subroutine based on Hudson's algorithm but slightly different from the one described in section 2.4.1a. When N_1 is known an iterative search routine as described in chapter 1 and again briefly in section 2.2.2 is performed in order to arrive at an estimate of the join point.

2.4.2 Calculation of Empirical Moments

The empirical moments calculated were the mean and variance over all the simulated samples. In order to obtain these moments the entire process described in section 2.4.1 was repeated 1000 times. It was felt that 1000 samples was a large enough number to establish the empirical distribution. The parameter estimates from the 1000 trials were then averaged and the variance in the estimates over all trials was calculated.

2.4.3 Estimation of the Asymptotic Moments

The asymptotic moments which were found in section 2.3.1 are in terms of the true parameter values. In the regression situation these values are unknown. Hence, it is customary to substitute the parameter estimates into the equations in order to obtain estimates of the expected values and variances or mean squared errors.

For each of the simulated samples, the regression coefficient estimates were substituted into the asymptotic formulas for the expected value (2.4) and the mean squared error (2.5). The average expected values and mean squared errors were then taken over all samples. These averages could then be compared to the sample means and variances of the parameters over all models to compare the numerical accuracy of the estimated asymptotic results and the empirical results.

Ideally, one would like to substitute estimates of the error variances as well since these parameters are also unknown. However, since there is no replication of the data, it is impossible to obtain reasonable estimates for the error variances. Hence, in this study the true error variances specified in the simulation were used in the estimation of the asymptotic moments.

2.5 RESULTS

A numerical model is used to illustrate the procedures described in this chapter. First the specific model and the resulting parameter values are given in section 2.5.1. The first comparison is made in section 2.5.2 between results from samples of two different sizes. Next, section 2.5.3 follows with a discussion of the expected value as a function of the measurement error variance for this particular model. The fourth section (2.5.4) presents results of the simulation study of the empirical distribution and compares these results when N_1 is known with those of the asymptotic results. A comparison is also made between results for different values of N_1 and N_1 unknown.

2.5.1 Numerical Model and Parameter Values

The model under study has a Type I join and is of the form

$$Y_i = \begin{cases} 23.38 - 1.73X_i\,, & X_i \leq 13.5144 \\ \\ -23.38 + 1.73X_i\,, & X_i \geq 13.5144 \end{cases}$$

The two segments were arbitrarily chosen to be symmetric with an angle of 60° between them. By varying the sample size, error variances and prior knowledge about X_{N_1}, whether or not N_1 is known, several kinds of comparisons can be made. The following table (Table 2.1) summarizes the combinations to be studied here.

Table 2.1 Parameter Values for Numerical Example

Parameters	Case 1	Case 2
N	30	60
N_1	13	26
N_2	17	34
$\bar{X}_1$	7.00	00.50
$s^2_{X(1)}$	15.17	58.50
$\bar{X}_2$	22.00	30.50
$s^2_{X(2)}$	25.50	99.17
σ^2_ε	.01, 0.1, 0.5, 1.0, 5.0	same as Case 1
σ^2_ν	0.0, 0.1, 0.5, 1.0, 5.0	same as Case 1 plus 0.01 and 0.05
$\underset{\sim}{X}$	(1,2,...30)	(-12,-11,...,47)
prior knowledge of X_{N_1}	13.5144	12.0, 13.0, 13.5144, 14.0, none

2.5.2 Comparison of Results for Two Sample Sizes

Since the asymptotic results of section 2.3 are based on large sample approximations, it is of interest to see how much better the approximation is when the sample size is doubled. Notice that not only is the total sample size N doubled from 30 to 60, the new X-values were chosen so that N_1 and N_2, the sample sizes in each segment, doubled as well.

The asymptotic expected values and MSE's of the regression parameters and join point for Case 1 when N = 30 and σ_ε^2 = .01 are given in Table 2.2. The entries are calculated from the true parameter values as if they were all known. Note that when σ_ν^2 = 0.0, the expected values are equal to the known values and therefore are unbiased since there is no measurement error. In addition, the variances are small. As σ_ν^2 increases the expected values of the regression parameters are further from the true value since the absolute bias increases. As the bias increases the MSE increases as well.

There is no change in the expected value with increasing values of σ_ε^2 since the Hudson technique is based on the least squares procedure which minimizes the sum of squares of the residual y_i. However, the MSE continues to increase in proportion to the increase in σ_ε^2.

Table 2.3 has the resulting statistics for Case 2 where N = 60. The expected values are, indeed, closer to the true values and, as expected, the MSE are smaller than the MSE based on the smaller sample size of N = 30.

An interesting difference is observed in Table 2.3 in the decrease of MSE values for the β_{10} intercept term. While the MSE of β_{20} decreases by about one third, the MSE for β_{10} decreases by a factor of 10. This result is most likely an artifact of the data due to the relatively small value for $\bar{X}_1$.

2.5.3 Expected Value as a Function of Measurement Error Variance

Note that the absolute bias of the regression functions is in each case steadily increasing as the measurement error increases. On the other hand, the expected value of the join point is not a monotonic function of the measurement error. By substituting the parameters from Case 2 into equations (2.65), (2.68), and (2.69) the asymptotic join point can be studied as a function of the measurement error. It can be seen that for $\sigma_\nu^2 = 0$ and $\sigma_\nu^2 = 74.79$, $\hat{\alpha}$ is an unbiased estimate of α; however, the MSE in the latter case would be considerably larger. From equations (2.67) and (2.68), it can be seen that the estimate of $\hat{\alpha}$ for this particular model reaches a maximum at $\sigma_\nu^2 = 30.95$. While these values of σ_ν^2 are of theoretical interest in studying the behavior of $\hat{\alpha}$ as a function of σ_ν^2, they have little practical value since it is unlikely that data with such large measurement error variances would be useful.

2.5.4 Empirical Distribution for Sample of Size N = 60

In order to study the empirical distribution of the join point the numerical model described in section 2.5.1 with sample size 60 was simulated according to the procedure outlined in section 2.4.1. This entire process was repeated 1000 times to produce 1000 different samples. The parameter estimates from the 1000 trials were then averaged and the variance in the parameter estimates over all 1000 trials was calculated. This was done for each of the combination of values for σ_ε^2 and σ_ν^2 and the various values of prior knowledge of X_{N_1} as well as for the case when N_1 is not specified. In addition, the parameter estimates were substituted into the equations for the asymptotic values to obtain estimates of the asymptotic values.

2.5.4a Case When N_1 is Known. In order to study the behavior of the estimation procedure for various values of N_1, the simulation was done separately for each value specified. The first value of x_{N_1} which will be studied is $x_{N_1} = 13.5144 = \alpha$.

A brief discussion of the choice of 13.5144 for the cutoff point of the first segment is in order here. The true X-values are the whole numbers (-12,-11,...,47) and hence the true value for X_{N_1} is 13.0, where $N_1 = 26$. However, when measurement error is introduced, x_{N_1} can be less than, equal to, or greater than 13.0 and possibly even greater than 13.5144 -- the true join point value. Hence, the true join point value was chosen in order to minimize the number of observed values which are misclassified.

Several points should be observable from the results of the simulation experiment. First, when $\sigma_\nu^2 = 0$ the mean of the join point estimates over all the samples should be equal to the true value of α. Second, the variance of the sample join point estimate should be close to the approximate MSE of $\hat{\alpha}$ calculated from the Taylor series expansion of equation (2.3). A third comparison can be made between the asymptotic values and the estimated asymptotic values based on the parameter estimates.

In the event that $\sigma_\nu^2 > 0$, and measurement error exists, the average of the join point estimates would not be expected to equal the true value since the estimates are known to be biased. This average should, however, compare favorably with the Taylorized expected value because this approximation includes the bias term also. In addition, the variance of the join point estimates should be close to the Taylor Series approximation to the MSE $(\hat{\alpha})$.

The simulated results for Case 2 where N = 60 are presented in Table 2. and are comparable to the asymptotic results of Table 2.3. The first point to be considered is the mean of the join point estimates over all 1000 samples when $\sigma_{\nu}^2 = 0$. From columns (1) and (2) of Table 2.4 it can be seen that, as was expected, this mean is within three standard errors of the true value in every case except for the last when $\sigma_{\varepsilon}^2 = 5.0$

Second, the variance of the sample join points should be close to the approximate MSE calculated from the Taylor Series expansion. Comparing column (3) of Table 2.4 and column (10) of Table 2.3 it can be seen that, while the two are not identical, they are similar in many cases and for the most part the amount of increase with increasing error is similar. Comparing the average estimates of the asymptotic moments (Table 2.4, col. 5) with the approximate asymptotic moments (Table 2.3, col. 10) shows an even more striking similarity.

The last point to be considered here is that the average of the join point estimates when $\sigma_{\nu}^2 > 0$ should be equal to the Taylorized expected values from Table 2.4 (col. 9). Looking at Table 2.4 (col. 1) it can be seen that none of the values are within three standard errors of the Taylorized expected values. In fact, the means from the simulated samples decrease as σ_{ν}^2 increases while the asymptotic values increase. Unlike the estimated asymptotic variances, the estimated asymptotic expected values are not similar to the asymptotic values, nor do they increase or decrease in any pattern.

A further point of interest when considering the case of N_1 known is the effect of the misspecification of N_1 or X_{N_1}, in this case. Table 2.5 presents the statistics describing the join point for the case when $N = 60$ and σ_{ε}^2 is set at 0.01. Note that the third set of entries for $X_{N_1} = 13.5144$ are the same as Table 2.4 for $\sigma_{\varepsilon}^2 = 0.01$.

When X_{N_1} is thought to be 13 -- which is actually X_{N_1} for this model -- there is very good agreement between the simulated and asymptotic results for the case of no measurement error. On the other hand, when there is measurement error the means and variances over the 1000 samples are not very close to the asymptotic values or even the approximate asymptotic moments. It is interesting to note, however, that the estimated asymptotic MSE is reasonably close to the asymptotic MSE for all values of σ_ν^2 .

When X_{N_1} is actually misspecified and thought to equal 12 or 14 the join point estimates are way out of line as can be seen in the first and fourth sets of entries of Table 2.5, col. 1. The effects of misspecifying X_{N_1} are especially noticeable in the case of no measurement error. Interestingly, the estimated asymptotic MSE's are close to the approximated values based on the Taylorized MSE.

The final set of entires in Table 2.5 is for the case when X_{N_1} is not specified at all. This case will be discussed further in section 2.5.4b, but it is interesting to note that for all values of σ_ν^2 the mean join point estimate is within three standard errors of the asymptotic value based on N_1 known. In fact, it is within three standard errors of the true join point no matter what the value of σ_ν^2 is.

2.5.4b <u>Case When N_1 Is Unknown</u>. Since knowledge about the value of N_1 is not always available, it is of interest to simulate the iterative procedure which can be used to fit the model. Even though the asymptotic results do not hold under these circumstances it is possible to compare the mean join point estimate over the 1000 samples with the true value of the join point itself.

It appears that for most of the values of σ_ε^2 and σ_ν^2 presented here that the mean join point estimate is within three standard errors of the true join point value. Looking at Table 2.6 a steady increase in the variance can be seen as both σ_ε^2 and σ_ν^2 increase, The two exceptions are for $\sigma_\varepsilon^2 = 1.0$, $\sigma_\nu^2 = 5.0$ and $\sigma_\varepsilon^2 = 5.0$, $\sigma_\nu^2 = 1.0$. On the other hand, there seems to be no particular pattern associated with the mean values.

Results not presented here indicate that the regression coefficients are not, in general, close to the true values. Perhaps since the join point is based on a ratio of the biased regression coefficients some of the bias cancels out to produce a more nearly consistent join point estimate.

2.6 CONCLUSIONS

It appears from the results presented in Section 2.5.4 that the simulation procedure does not behave quite as expected based on the asymptotic results. First, when N_1 is considered to be known, not all of the variances of the computed join point agree with the estimated asymptotic MSE or the asymptotic MSE (see Table 2.4, col. 3). Second, the averages of the join point estimates do not follow the same pattern as the asymptotic expected values as σ_ν^2 changes, and they are not close to either the asymptotic values or the true values.

There are several possible explanations for these differences. One such explanation for the differences could be in the form of sampling error attributed to the random number generator which produces the σ_ε^2 and σ_ν^2. It is very likely with samples of size $N = 60$ that not every set of 60 random normal deviates has a mean of 0.0 and the exact variance specified. Thus parameter estimates based on such samples would have additional forms of error not included in the theoretical results.

In addition to sampling errors, several levels of approximation have been used to obtain the asymptotic results. For instance, the asymptotic results are only approximations to begin with because the Taylor series itself is truncated after the second order term. In addition, the variances and covariances of the regression coefficients which determine the join point are based on large sample assumptions which are perhaps not met by samples of size $N_1 = 27$ and $N_2 = 34$.

A further consideration is the relationship between the measurement error and the distance between the X-values. It was stated earlier that X_{N_1} was chosen to be 13.5.44 in order to minimize the number of observed data points classified in the wrong segment. If the maximum error in measurement is small enough, say less than half the distance between two data points, then it will be unlikely that the observed values will be misclassified.

It appears that in actual practice it does more harm than good to the estimation results to specify the value of X_{N_1} when measurement error exists unless one is fairly certain where the true join point lies. The results in Table 2.5 show that the results of misspecifying the value by only one data point are quite large. Even specifying the value as $X_{N_1} = 13.0$, the value based on the true X-values, leads to erroneous results.

Based on this study, it would seem preferable to use the iterative procedure in order to obtain the estimates -- especially if one is interested in the join point estimate and not the regression coefficients -- rather than risk specifying the wrong value for X_{N_1}. However, it is to be hoped that the alternative estimation procedure set forth in chapter 3 will prove to be better for estimating both the join point and the regression coefficients when measurement error exists.

Table 2.2

Asymptotic Expected Values and MSE Estimates for Regression Coefficients and Join Point Based on the True Parameter Values for Various Values of σ_ε^2 and σ_ν^2 When Estimates Are Based on the Hudson Technique and the Sample Size Is N=30

σ_ε^2	σ_ν^2	$E(\beta_{10})$[1] (1)	MSE (2)	$E(\beta_{11})$[1] (3)	MSE (4)	$E(\beta_{20})$[1] (5)	MSE (6)	$E(\beta_{21})$[1] (7)	MSE (8)	$E(\alpha)$[1] (9)	MSE (10)
.01	0.0	23.3800	.003254	-1.7300	.000051	-23.3800	.011753	1.7300	.000023	13.5144	.000432
	0.1	23.3007	.077143	-1.7187	.001558	-23.2313	.344564	1.7232	.000711	13.5192	.068645
	0.5	22.9935	.363604	-1.6748	.007397	-22.6481	1.650320	1.6967	.003408	13.5374	.349363
	1.0	22.6309	.702866	-1.6230	.014288	-21.9438	3.227430	1.6647	.006666	13.5580	.717896
	5.0	20.3775	2.892280	-1.3011	.057118	-17.1406	14.010900	1.4464	.028880	13.6555	4.361190
0.1	0.0		.032544		.000507		.117532		.000231		.004319
	0.1		.106287		.002012		.449950		.000917		.072559
	0.5		.392181		.007839		1.754160		.003612		.353382
	1.0		.730773		.014716		3.329420		.006865		.722047
	5.0		2.916020		.057462		14.100200		.029054		4.366420
0.5	0.0		.162722		.002536		.587659		.001153		.021596
	0.1		.235813		.004027		.918332		.001837		.089953
	0.5		.519185		.009803		2.215700		.004517		.371245
	1.0		.854801		.016620		3.782690		.007753		.740498
	5.0		3.021550		.058987		14.497100		.029825		4.389680
1.0	0.0		.325444		.005072		1.175320		.002307		.043193
	0.1		.397721		.006547		1.503810		.002986		.111695
	0.5		.677942		.012258		2.792620		.005648		.393573
	1.0		1.009840		.018999		4.349280		.008863		.763561
	5.0		3.153460		.060894		14.993200		.030790		4.418760
5.0	0.0		1.627220		.025359		5.876590		.011534		.215964
	0.1		1.692980		.026701		6.187630		.012177		.285635
	0.5		1.947990		.031898		7.408010		.014698		.572198
	1.0		2.250120		.038031		8.882020		.017742		.948065
	5.0		4.208770		.076152		18.962400		.038504		4.651390

[1] Asymptotic expected values are the same for all values of σ_ε^2 so they are given only for $\sigma_\varepsilon^2 = 0.01$.

Table 2.3

Asymptotic Expected Values and Mean Squared Error Estimates for Regression Coefficients and Join Point Based on the True Parameter Values for Various Values of σ_ε^2 and σ_ν^2 When Estimates Are Based on the Hudson Technique and the Sample Size Is N=60

σ_ε^2	σ_ν^2	$E(\hat\beta_{10})$	MSE	$E(\hat\beta_{11})$	MSE	$E(\hat\beta_{20})$	MSE	$E(\hat\beta_{21})$	MSE	$E(\alpha)$	MSE
0.01	0.00	23.3800	.000386	-1.7300	.000007	-23.3800	.003053	1.7300	.000003	13.5144	.000221
	0.01	23.3798	.000391	-1.7297	.000026	-23.3747	.011310	1.7298	.000012	13.5147	.001347
	0.05	23.3793	.000411	-1.7285	.000105	-23.3534	.044318	1.7291	.000047	13.5157	.005855
	0.10	23.3785	.000436	-1.7270	.000203	-23.3268	.085542	1.7283	.000092	13.5170	.011498
	0.50	23.3727	.000653	-1.7153	.000982	-23.1153	.413849	1.7213	.000445	13.5271	.056984
	1.00	23.3655	.000977	-1.7009	.001941	-22.8532	.820564	1.7127	.000882	13.5394	.114691
	5.00	23.3119	.007964	-1.5937	.009070	-20.8473	3.935110	1.6470	.004228	13.6263	.610315
0.10	0.00		.003863		.000066		.030531		.000030		.002212
	0.01		.003868		.000085		.038785		.000039		.003338
	0.05		.003887		.000164		.071784		.000074		.007847
	0.10		.003912		.000262		.112995		.000118		.013492
	0.50		.004129		.001041		.441203		.000471		.058992
	1.00		.004529		.001999		.847794		.000908		.116716
	5.00		.011439		.009124		3.961400		.004254		.612479
0.50	0.00		.019313		.000329		.152657		.000148		.011060
	0.01		.019318		.000348		.160900		.000157		.012187
	0.05		.019338		.000427		.193854		.000193		.016703
	0.10		.019363		.000525		.235010		.000237		.022356
	0.50		.019579		.001301		.562775		.000589		.067917
	1.00		.019902		.002258		.968818		.001026		.125718
	5.00		.026884		.009367		4.078230		.004366		.622101
1.00	0.00		.038626		.000657		.305314		.000297		.022119
	0.01		.038631		.000677		.313543		.000305		.023249
	0.05		.038650		.000755		.346441		.000341		.027772
	0.10		.038676		.000853		.387528		.000385		.033434
	0.50		.038891		.001627		.714740		.000737		.079073
	1.00		.039214		.002581		1.120100		.001172		.136970
	5.00		.046190		.009670		4.224260		.004508		.634128
5.00	0.00		.193130		.003287		1.526570		.001483		.110595
	0.01		.193134		.003306		1.534690		.001492		.111740
	0.05		.193154		.003383		1.567140		.001527		.116325
	0.10		.193178		.003478		1.607670		.001570		.122065
	0.50		.193389		.004235		1.930460		.001917		.168320
	1.00		.193706		.005167		2.330340		.002347		.226988
	5.00		.200642		.012092		5.392540		.005637		.730345

Table 2.4

Statistics Describing the Distribution of the Join Point Over 1000 Samples of Size N=60 for Various Values of σ_ε^2 and σ_ν^2 When Prior Knowledge Specifies x_{N_1}=13.5144 and the Estimates Are Based on the Hudson Technique

σ_ε^2	σ_ν^2	Mean Over All Samples (1)	Standard Error of the Mean (2)	Variance of Computed Join Points (3)	Mean of Asymptotic Expected Values (4)	Mean of Asymptotic Variances (5)
0.01	0.00	13.5143	.000475	.000225	13.5235	.000221
	0.01	13.5129	.000987	.000974	13.5334	.001349
	0.05	13.5098	.003023	.009140	13.5315	.005860
	0.10	13.5046	.003601	.012965	13.5183	.011545
	0.50	13.4623	.009327	.086994	13.4948	.057681
	1.00	13.4465	.011575	.133977	13.4254	.116166
	5.00	13.3841	.018520	.342971	13.6991	.648516
0.10	0.00	13.5138	.001485	.002205	13.4991	.002196
	0.01	13.5127	.001736	.002998	13.5526	.003340
	0.05	13.5045	.003450	.011906	13.5004	.007887
	0.10	13.4977	.005772	.033320	13.4948	.013528
	0.50	13.4822	.008136	.066192	13.4161	.058437
	1.00	13.4442	.011808	.139433	13.6881	.120776
	5.00	13.3541	.020524	.421231	13.5265	.653173
0.50	0.00	13.5163	.003338	.011144	13.6362	.010960
	0.01	13.5117	.003428	.011750	13.5871	.012498
	0.05	13.4962	.005901	.034825	13.5115	.016588
	0.10	13.4789	.006997	.048957	13.6598	.022175
	0.50	13.4585	.010458	.109378	13.4621	.067986
	1.00	13.4333	.012447	.154929	13.5795	.128918
	5.00	13.3591	.019561	.382618	13.2478	.626175
1.00	0.00	13.5135	.004750	.022559	13.6048	.022084
	0.01	13.5155	.004829	.023323	13.5560	.023690
	0.05	13.4904	.005827	.035792	13.5754	.027287
	0.10	13.4826	.006800	.046238	13.4585	.034470
	0.50	13.4332	.011050	.122110	13.6106	.082484
	1.00	13.4263	.012845	.164995	13.3381	.138261
	5.00	13.3789	.018786	.352931	13.8602	.688213
5.00	0.00	13.4377	.009169	.084072	13.4081	.105136
	0.01	13.4181	.009518	.090592	13.4201	.119500
	0.05	13.4214	.010441	.109018	13.1219	.109920
	0.10	13.4016	.012274	.150638	11.4656	.117999
	0.50	13.3929	.014481	.209698	13.3953	.164442
	1.00	13.3761	.015836	.250789	13.1221	.222044
	5.00	13.3018	.021595	.466341	13.5316	.789734

Table 2.5

Statistics Describing the Empirical Distribution of the Join Point for Various Values of σ_ν^2 and the Prior Knowledge About N_1 When σ_ε^2=.01 and Estimates Are Based on the Hudson Technique

x_{N_1}	σ_ν^2	Mean Over All Samples (1)	Standard Error (2)	Variance Over All Samples (3)	MSE[1] (4)	Mean Over All Samples of Estimated Asymptotic: Expected Values (5)	Mean Over All Samples of Estimated Asymptotic: MSE (6)	Asymptotic Expected Value[2] (7)	MSE[2] (8)
12.00	0.00	12.0000	.000000	.000000	2.2934	12.0000	.000211	13.5144	.000221
	0.01	11.4655	.014726	.216840	4.4161	11.0614	.001012	13.5147	.001347
	0.05	11.4279	.015912	.253194	4.6129	10.7813	.004119	13.5159	.005855
	0.10	11.4418	.019265	.371133	4.6776	10.6920	.007961	13.5170	.011498
	0.50	11.4621	.029283	.857471	5.1217	10.9095	.040644	13.5271	.056984
	1.00	11.5631	.032768	1.073760	4.9795	10.6345	.080088	13.5394	.114691
	5.00	11.6813	.035579	1.265860	5.0489	13.8649	.621937	13.6263	.610315
13.00	0.00	13.5143	.000475	.000225	.0002	13.5235	.000221		
	0.01	12.7833	.023866	.569584	1.1045	13.5335	.001330		
	0.05	12.7831	.024185	.584922	1.1219	12.1814	.004987		
	0.10	12.8278	.024108	.581208	1.0562	12.2574	.009860		
	0.50	12.9227	.024010	.576459	.9418	13.4620	.055940		
	1.00	12.9104	.024736	.611881	1.0075	12.5903	.103512		
	5.00	12.9017	.027340	.747472	1.2725	13.7542	.637762		
13.5144	0.00	13.5143	.000475	.000225	.0003	13.5235	.000221		
	0.01	13.5129	.000987	.000974	.0010	13.5334	.001349		
	0.05	13.5098	.003023	.009140	.0092	13.5315	.005860		
	0.10	13.5046	.003601	.012965	.0131	13.5183	.011545		
	0.50	13.4623	.009327	.086994	.0912	13.4948	.057681		
	1.00	13.4465	.011575	.133977	.1426	13.4254	.116166		
	5.00	13.3841	.018520	.342971	.4016	13.6991	.648516		
14.00	0.00	14.0000	.000000	.000000	.2358	14.0000	.000221		
	0.01	13.7151	.006528	.042621	.0828	13.5333	.001382		
	0.05	13.6801	.005810	.033758	.0607	13.5139	.006044		
	0.10	13.6589	.005552	.030830	.0510	13.5173	.011889		
	0.50	13.6453	.006331	.040084	.0541	13.7265	.060844		
	1.00	13.6610	.008061	.064974	.0798	13.4155	.119438		
	5.00	13.7057	.014758	.217796	.2241	13.6482	.663345		

Table 2.5, continued

x_{N_1}	σ_ν^2	Mean Over All Samples (1)	Standard Error (2)	Variance Over All Samples (3)	MSE (4)	Mean Over All Sample of Estimated Asymptotic: Expected Values (5)	Mean Over All Sample of Estimated Asymptotic: MSE (6)	Asymptotic Expected Value[2] (7)	MSE[2] (8)
none[3]	0.00	13.5143	.000475	.000225					
	0.01	13.5129	.000987	.000974					
	0.05	13.5144	.001977	.003908					
	0.10	13.5125	.002799	.007836					
	0.50	13.5232	.006215	.038628					
	1.00	13.5403	.009069	.082249					
	5.00	13.6164	.037265	1.388660					

[1] $MSE = col(3) + [col(7) - col(1)]^2$

[2] Asymptotic values not applicable

[3] Asymptotic values are the same for all values of x_{N_1}

Table 2.6

Empirical Moments of the Join Point Estimates Based on the Hudson Iterative Procedure for Various Values of σ_ε^2 and σ_ν^2 When There is No Prior Knowledge About N_1

σ_ε^2	σ_ν^2	Mean Over All Samples (1)	Standard Error (2)	Variance Over All Samples (3)
0.01	0.00	13.5143	.000475	.000225
	0.01	13.5129	.000987	.000974
	0.05	13.5144	.001977	.003908
	0.10	13.5125	.002799	.007836
	0.50	13.5232	.006215	.038628
	1.00	13.5403	.009069	.082249
	5.00	13.6164	.037265	1.388660
0.10	0.00	13.5141	.001502	.002255
	0.01	13.5127	.001732	.002998
	0.05	13.5119	.002412	.005818
	0.10	13.5121	.003140	.009861
	0.50	13.5225	.006472	.041884
	1.00	13.5384	.009254	.085644
	5.00	13.6235	.039106	1.529320
0.50	0.00	13.5137	.003358	.011277
	0.01	13.5117	.003430	.011765
	0.05	13.5174	.004171	.017400
	0.10	13.5111	.004383	.019206
	0.50	13.5202	.007323	.053628
	1.00	13.5111	.027120	.735524
	5.00	13.6246	.037358	1.395600
1.00	0.00	13.5136	.004762	.022672
	0.01	13.5120	.004862	.023638
	0.05	13.5098	.005091	.025916
	0.10	13.5105	.005542	.030714
	0.50	13.5194	.008213	.067447
	1.00	13.5090	.027766	.770939
	5.00	13.5965	.020838	.434240
5.00	0.00	13.5106	.011225	.125997
	0.01	13.5056	.011069	.122528
	0.05	13.5464	.054416	1.184430
	0.10	13.5409	.034530	1.192300
	0.50	13.5597	.067620	4.572530
	1.00	13.5118	.047787	2.283480
	5.00	13.6487	.040429	1.634530

CHAPTER 3

WALD'S GROUPING TECHNIQUE APPLIED TO PIECEWISE LINEAR REGRESSION

3.1 INTRODUCTION

In the usual regression situation measurement error in the independent variable leads to biased regression coefficients when OLS estimation procedures are employed. It was seen in the previous chapter that in a PWLR model the join point estimate is also biased when OLS procedures are employed.

Wald (1940) proposed a grouping technique which, when applied to data in which measurement error exists, produces asymptotically unbiased estimates of the regression coefficients. Other authors, among them Bartlett (1949), proposed variations of the grouping techniques, all of which can be applied to data which have a single linear relation between the dependent and independent variables.

It is the purpose here to study the use of grouping techniques when the model is a PWLR model rather than a single straight line. In addition, a technique is proposed for estimating the variance numerically from the sample data. When the Wald estimate is used, not only is the join point estimate based on a ratio, but the slope estimates are ratios as well. Thus, the distribution of $\hat{\alpha}$ is complicated and the usual Taylor series expansions have proven unwieldy in approximating the variances of the join point estimate based on grouping techniques. Dorff and Gurland (1961a) give asymptotic variances for regression slopes for the Wald-type estimates. However, the join point remains a problem. An approach proposed by Tepping (1968) will be used to obtain variance estimates for all the parameters. This approach is based on the Taylor series expansion, but it uses actual numerical values to obtain the variance estimates rather than obtaining a closed form mathematical solution.

In order to study the feasibility of using a Wald-type estimator in the PWLR model, the following approach is taken. The additional assumption necessary for this technique when applied to the model presented in section 2.2.1 is explained in section 3.2 along with an outline of the estimation procedure. Next, some of the problems encountered in applying the procedure and solutions to these problems are discussed in section 3.3. Section 3.4 briefly discusses the asymptotic expected value of the estimator. In section 3.5 an outline of the general approach for the variance estimation problem is given along with the specific application to the Wald estimator. A simulation procedure parallel to that described in section 2.4 is discussed in section 3.6. Finally, the techniques are applied to a numerical model and the results are seen in section 3.7.

3.2 MODEL AND ESTIMATION PROCEDURE

3.2.1 Additional Assumptions for the Model

The model used in this aspect of the study is the same as the model described in section 2.2.1 with the additional assumption that

$$\liminf_{N_1 \to \infty} \left| (\Sigma_{11} p^{-1} x_i - \Sigma_{12} p^{-1} x_i)/N_1 \right| > 0$$

$$\liminf_{N_1 \to \infty} \left| (\Sigma_{21} p^{-1} x_i - \Sigma_{22} p^{-1} x_i)/N_2 \right| > 0 \qquad (3.1)$$

where p represents the proportion of observations in each group G_{ij}, and Σ_{ij} denotes summation over observations in the ij-th group. These groups, G_{ij}, were first mentioned in chapter 1, section 1.3.2 and are all defined more explicitly in section 3.2.2. This assumption is necessary to insure that the denominators of the slope estimates which are given in 3.2.2 do not equal 0. The manner of assignment of the observations to groups will be discussed later in section 3.3.1. It suffices here to say that the subdivision of the observations into groups must be independent of the errors ε and ν.

3.2.2 Estimation Procedures

In order to clarify the Wald-type estimators in the PWLR we introduce the following additional notation.

First, the observed data points are ordered as follows: $x_1 < \ldots < x_{N_1} < x_{N_1+1} < \ldots < x_N$, where it is known that α lies in the interval $[x_{N_1}, x_{N_1+1}]$. There are then two simple linear segments such that the first pN_1 observations from the first segment form G_{11}, the first group, of the first segment, and the last pN_1, observations from the first segment form G_{12}. If $2p < 1$ then the middle group of observations is discarded. Similarly, the first pN_2 and last pN_2 observations of the second segment form groups G_{21} and G_{22}.

Wald (1940) assumes $p = \frac{1}{2}$ and proposes the following consistent estimators for the regression coefficients of a single segment.

$$\tilde{\beta}_{11} = \frac{(\Sigma_{11}y_i - \Sigma_{12}y_i)/N_1}{(\Sigma_{11}x_i - \Sigma_{12}x_i)/N_1} \qquad \tilde{\beta}_{21} = \frac{(\Sigma_{21}y_i - \Sigma_{22}y_i)/N_2}{(\Sigma_{21}x_i - \Sigma_{22}x_i)/N_2} \tag{3.2}$$

$$\tilde{\beta}_{10} = \bar{y}_1 - \tilde{\beta}_{11}\bar{x}_1 \qquad \tilde{\beta}_{20} = \bar{y}_2 - \tilde{\beta}_{21}\bar{x}_2 \tag{3.3}$$

where $\bar{y}_k$ and $\bar{x}_k$ are as defined in section 2.3. Bartlett (1947) proposed using $p = 1/3$ rather than $p = 1/2$, but the resulting estimators are the same; only the groups are different. Several studies (Thiel and Yzeren, 1956) showed that the efficiency of $\tilde{\beta}$, compared to $\hat{\beta}$ is maximum when p was near 1/3. However, for the rest of this study p is taken to be 1/2.

Taking these estimates of the regression coefficients for whatever value of p, the join point estimate is

$$\tilde{\alpha} = \frac{\tilde{\beta}_{10} - \tilde{\beta}_{20}}{\tilde{\beta}_{21} - \tilde{\beta}_{11}} \tag{3.4}$$

If the estimate $\tilde{\alpha}$ does not fall in the interval $[x_i, x_{i+1}]$ then an additional constraint must be imposed so that the join point estimate corresponds to one of the data points.

The constrained slope estimators are

$$\tilde{\beta}^*_{11} = \frac{[\Sigma_{11}y_i/(N_1/2)] - y_{N_1}}{[\Sigma_{11}x_i/(N_1/2)] - x_{N_1}} \quad \text{and} \quad \tilde{\beta}^*_{21} = \frac{y_{N_1} - [\Sigma_{22}y_i/(N_2/2)]}{x_{N_1} - [\Sigma_{22}x_i/(N_2/2)]}$$

so that $\tilde{\alpha}^* = x_{N_1}$.

For the case when there is no prior knowledge of the location of the join point, the iterative procedure is parallel to that proposed by Hudson and described in section 2.2.2 except that the Wald estimators were used for the regression coefficient estimates rather than the ordinary least squares estimates.

3.3 SOME PROBLEMS IN APPLYING A GROUPING TECHNIQUE AND PROPOSED SOLUTIONS

3.3.1 Grouping the Observations

A practical problem in applying any one of the grouping techniques to any linear regression problem is that of dividing the observed values into groups. The basic assumption of the technique is that the method of grouping be independent of the error terms ε and ν.

Random assignment to groups would lead to a violation of assumption (vii). Hence we turn to a second possible method of grouping which involves ordering the data pairs (x_i,y_i) on x_i, finding the desired percentiles, and then dividing the data into two equal groups depending upon whether x_i was below the lowest percentile or above the highest. In the case of Wald (1940), the percentile would be the median $x_{(1/2)}$ and for Bartlett (1947) the tertiles, $x_{(1/3)}$, $x_{(2/3)}$, would define the groups. Other possibilities involve the use of quartiles.

However, this method of grouping depends on the observed x_i. Therefore, in general, it is not independent of the errors $\underset{\sim}{\nu}$. If it is known a priori that the x_i have the same ordering as the X_i then this grouping method could be considered independent.

In addition, the groupings will be independent if $\max|\nu_i| < \frac{1}{2}\min|X_i-X_j|$, $(i \neq j)$, which requires the errors to be very small or the intervals between X-values to be large. According to Madansky (1959), one might known that there exists a δ such that $P\{|\nu_i| \geq \delta\}$ is negligibly small, and that the number of x_i's in the interval $[x_p-\delta, x_p+\delta]$, where x_p is the p'th percentile, is also very small. In this case, even though the grouping is not strictly independent of the ν, there is a high probability that the grouping will insure that the order of the x_i's is the same as the order of the X_i's.

The number δ may be determined by some a priori knowledge about the possible range of ν_i. For example, if one assumes that the errors are normally distributed then one can choose $\delta = 4\sigma_\nu$, say, so that $P\{|\nu| \geq 4\sigma_\nu\} = 0.0002$. If δ represents the maximum value that the error is likely to take, then 2δ is the minimum interval between any two X-values which will insure a high probability of independence of grouping and error. If the distance between X-values is less than 2δ then the consistency of the slope estimate is questionable. It can be seen that 2δ becomes large rapidly and is already over 2 units when σ_ν^2 is only 0.10. Thus, unless the distribution of the X-values is widely spread, the slope estimate is likely to be biased because the assumption of grouping independent of error is violated. For various values of σ_ν^2, we find the following values for δ and 2δ.

Table 3.1 Maximum Value of Error and Minimum Interval Between Abscissae

σ_ν^2	δ	2δ
0.01	0.40	0.80
0.05	0.89	1.79
0.10	1.26	2.52
0.50	2.83	5.66
1.00	4.00	8.00
5.00	8.94	17.89

If the following assumption is added to the six already set forth in sections 2.2.1 and 3.2.1 then the grouping of the observations can be said to be independent of the errors since the ν_i are restricted to a finite range.

Assumption (viii)

$$PR\{|\nu_i| \geq c/2\} \simeq 0 \quad \text{where} \quad c = \min_i |X_{i+1} - X_i| \quad . \tag{3.5}$$

Hence, the ν_i are restricted to a finite range.

3.3.2 Knowledge About N_1

At this point, a second major drawback of using the grouping approach in estimating the regression parameters in a PWLR model is the fact that N_1 must be known. In the least squares procedure, discussed previously, N_1 known was necessary for the theoretical developments, but, iterative procedures could be used to find the estimates.

For the grouping technique no criteria have been proposed previously for determining the best fit in order to estimate the join point. In addition, any attempts to establish the location of the join will be complicated by the measurement error.

It should be noted that assumption (viii) which insures that the grouping is independent also ameliorates the problem of a priori knowledge about the location of the join since the ordering of the x_i is the same as the true X_i. If 2δ is the minimum distance between any two X-values, then there is a high probability that the observed x_{N_1} corresponds to X_{N_1} no matter what the estimation procedure.

In the case where N_1 is unknown an iterative search procedure similar to that described in Section 2.2.2 will be used. The minimum residual sum of squares based on the Wald-type parameter estimates will be used to determine the model which provides the best fit.

3.4 ASYMPTOTIC EXPECTED VALUE

In section 2.3.1 it was shown that the join point estimate based on an OLS technique is biased when measurement error is present in the independent variable. As an alternative to OLS estimation, it is proposed that the Wald grouping technique be used in order to obtain the slope estimates which will, in turn, be used in estimating the join point. In this section, it is shown that the join point estimate based on Wald-type slope estimates is indeed a consistent estimate.

The consistency of the regression parameter estimates, $\tilde{\beta}_{10}$, $\tilde{\beta}_{11}$, $\tilde{\beta}_{20}$, and $\tilde{\beta}_{21}$ was proved by Wald (1940). Using the fact that the $\tilde{\beta}$'s are consistent and applying Slutsky's Theorem (Rao, 1952, p. 173) first to the differences of the slope estimates then to the differences of the intercept estimates and, finally, to the rate of these differences, it is easily shown that

$$\tilde{\alpha} \xrightarrow[N_1, N_2 \to \infty]{} \alpha$$

3.5 VARIANCE ESTIMATION PROCEDURE

When obtaining parameter estimates it is desirable to have a method of calculating the variance of the estimates as well. Variances for the slope estimates are given by Dorff and Gurland (1961a) as:

$$\mathrm{Var}(\tilde{\beta}_{11}) = \frac{N_1}{(\Sigma_{11}X_i - \Sigma_{12}X_i)^2} (\beta_{11}^2\sigma_\nu^2 + \sigma_\varepsilon^2)$$

and

$$\mathrm{Var}(\tilde{\beta}_{21}) = \frac{N_2}{(\Sigma_{21}X_i - \Sigma_{22}X_i)^2} (\beta_{21}^2\sigma_\nu^2 + \sigma_\varepsilon^2) \quad .$$

Unfortunately, these approximations contain the error variances which are not estimable unless the data are replicated.

In dealing with the problem of finding a variance approximation for the join point estimate based on the grouping technique one is faced with two problems. First, the usual approach using a Taylor series expansion is unwieldy. Second, if one could obtain a variance approximation by this method it would undoubtedly require the error variances. In order to circumvent the problem of lack of error variance and still obtain an approximation to the variance a numerical approach based on a Taylor series expansion is used.

A general outline of the numerical approach as proposed by Tepping (1968) is given in section 3.5.1. The application of this approach as applied to the current problem is described in section 3.5.2. Section 3.5.3 introduces the concept of pseudo replication , a further technique needed to arrive at the variance estimate.

3.5.1 General Description of Tepping Approach to Variance Estimation

In general, let $\underset{\sim}{u}$ be a k x 1 vector of sample statistics such that $\underset{\sim}{u}^T = (u_1 u_2 \ldots u_k)$ and let $\underset{\sim}{U}$ be the corresponding vector of population parameters such that $E(\underset{\sim}{u}) = \underset{\sim}{U}$. Take $f_j(u_1 u_2 \ldots u_k) = f_j(\underset{\sim}{u})$ to be a function of the k sample statistics and assume that $f_j(\underset{\sim}{u})$ estimates $f_j(\underset{\sim}{U})$. Let $f(\underset{\sim}{u})$ represent the p x 1 vector of functions $f_j(\underset{\sim}{x})$ which represent the $j = 1, \ldots p$ estimates of the parameters of interest.

By letting $f'(\underset{\sim}{U})$ be the p x k matrix of first partial derivatives of $f_j(\underset{\sim}{u})$ with respect to u_i evaluated at $\underset{\sim}{U}$, the first order Taylor series expansion (Kendall and Stuart, p. 231) of $f(\underset{\sim}{u})$ can be written in matrix notation as

$$\underset{\sim}{f}(\underset{\sim}{u}) = f(\underset{\sim}{U}) + f'(\underset{\sim}{U})[\underset{\sim}{u} - \underset{\sim}{U}] + O(n^{-1}) \quad . \tag{3.6}$$

For the case when p = 1, equation (3.6) is equivalent to

$$f(\underset{\sim}{u}) = f(\underset{\sim}{U}) + \sum_{i=1}^{k} \left[\frac{\partial f(\underset{\sim}{u})}{\partial u_i} \bigg|_{\underset{\sim}{U}} \right] (u_i - U_i) + O(n^{-1}) \quad . \tag{3.7}$$

Since $E(\underset{\sim}{u}-\underset{\sim}{U}) = 0$, it can easily be shown that $E(f(\underset{\sim}{u}) - f(\underset{\sim}{U})] = 0$ to the same order of approximation. Hence, the (p x p) variance-covariance matrix (VAR) of the matrix $f(\underset{\sim}{u})$ can be approximated by

$$VAR[f(\underset{\sim}{u})] = E\{[f'(\underset{\sim}{U})(\underset{\sim}{u} - \underset{\sim}{U})][f'(\underset{\sim}{U})(\underset{\sim}{u} - \underset{\sim}{U})]^T\} \tag{3.8}$$

For the sake of clarity, the following discussion will assume p = 1; however, it is not difficult to generalize p. The variance of $f(\underset{\sim}{u})$ when p = 1 is then:

$$\begin{aligned} Var[f(\underset{\sim}{u})] &= E\{[f(\underset{\sim}{u}) - f(\underset{\sim}{U})]^2\} \\ &= E\left\{ \left[\sum_{i=1}^{k} \left(\frac{\partial f(\underset{\sim}{u})}{\partial \underset{\sim}{u}_i} \bigg|_{\underset{\sim}{U}} \right) (\underset{\sim}{u}_i - \underset{\sim}{U}_i) \right]^2 \right\} \end{aligned} \tag{3.9}$$

Using this version of the Taylor series expansion for the variance is unwieldy for large values of k -- which is usually the case. Hence, Tepping (1968) proposed the following approach which uses the actual numerical values of the linearized sum (3.7).

Define a new statistic w such that

$$w = \sum_{i=1}^{k} (\underset{\sim}{u}_i - \underset{\sim}{U}_i) \left(\frac{\partial f(\underset{\sim}{u})}{\partial \underset{\sim}{u}_i} \bigg|_{\underset{\sim}{U}} \right) \tag{3.10}$$

and observe that $E(w) = 0$ since $\left(\frac{\partial f(\underset{\sim}{u})}{\partial \underset{\sim}{u}_i} \bigg|_{\underset{\sim}{U}} \right)$ is a constant. Now expression (3.9) can be rewritten as

$$\mathrm{Var}[f(\underset{\sim}{u})] = E(w^2) = \mathrm{Var}(w) \tag{3.11}$$

If Z is such that

$$Z = \sum_{i=1}^{k} \left(\left. \frac{\partial f(u)}{\partial u_i} \right|_{\underset{\sim}{U}} \right) u_i \quad , \tag{3.12}$$

then expression (3.10) becomes $w = Z - \sum_{i=1}^{k} \left(\left. \frac{\partial f(\underset{\sim}{u})}{\partial \underset{\sim}{u}_i} \right|_{\underset{\sim}{U}} \right) U_i$. Since the summation term is a constant, and u_i and u_j are uncorrelated for $i \neq j$, the variance of $f(\underset{\sim}{u})$ can be written as:

$$\begin{aligned} \mathrm{Var}[f(\underset{\sim}{u})] = \mathrm{Var}(Z) &= \sum_{i=1}^{k} \mathrm{Var}\left[\left(\left. \frac{\partial f(\underset{\sim}{u})}{\partial u_i} \right|_{\underset{\sim}{U}} \right) u_i \right] \\ &= \sum_{i=1}^{k} \mathrm{Var}(z_i) \end{aligned} \tag{3.13}$$

where $z_i = \left(\left. \frac{\partial f(\underset{\sim}{u})}{\partial u_i} \right|_{\underset{\sim}{U}} \right) u_i$.

A problem which can no longer be ignored is the fact that $\left(\left. \frac{\partial f(\underset{\sim}{u})}{\partial u_i} \right|_{\underset{\sim}{U}} \right)$ is a function of the unknown population parameters. The usual solution, and one which will be practiced here, is to substitute the sample estimates for the population parameters. For large enough sample sizes this procedure gives satisfactory results.

3.5.2 Approach Applied to Wald-Type Estimator

Turning now to the specific application at hand, let u_i be the observed pair of sample data points (x_i, y_i). Then

$$\underset{\sim}{u}^T = \begin{pmatrix} x_1 & \cdots & x_N \\ y_1 & \cdots & y_N \end{pmatrix}$$

and let the parameters of interest be represented by $f(\underset{\sim}{u})^T =$ $(\beta_{11}\beta_{10}\beta_{21}\beta_{20}\alpha)$ where the β_{ij} and α are estimated by (3.2), (3.3) and (3.4). Thus, $k = N$ and $p = 5$ from the general formulae presented earlie.

in this section. The matrix of partial derivatives of $f(\underset{\sim}{u})$ with respect to u_i are most easily presented in tabular form. Note that equation (3.4) has been rewritten as $f_5(u) = = \frac{A - B - C + D}{E - F}$ where

$$\begin{aligned} A &= N_2(\Sigma_{1.}y_i)(\Sigma_{11}x_i-\Sigma_{12}x_i)(\Sigma_{21}x_i-\Sigma_{22}x_i) \quad , \\ B &= N_2(\Sigma_{1.}x_i)(\Sigma_{11}y_i-\Sigma_{12}y_i)(\Sigma_{21}x_i-\Sigma_{22}x_i) \quad , \\ C &= N_1(\Sigma_{2.}y_i)(\Sigma_{21}x_i-\Sigma_{22}x_i)(\Sigma_{11}x_i-\Sigma_{12}x_i) \quad , \\ D &= N_1(\Sigma_{2.}x_i)(\Sigma_{21}y_i-\Sigma_{22}y_i)(\Sigma_{11}x_i-\Sigma_{12}x_i) \quad , \\ E &= N_1N_2(\Sigma_{21}y_i-\Sigma_{22}y_i)(\Sigma_{11}x_i-\Sigma_{12}x_i) \quad \text{and} \\ F &= N_1N_2(\Sigma_{11}y_i-\Sigma_{12}y_i)(\Sigma_{21}x_i-\Sigma_{22}x_i) \quad . \end{aligned} \tag{3.14}$$

Also, the derivative of any one of the functins $f_j(\underset{\sim}{u})$ depends on the group G_{ij} in which the element u_i is located and on whether the u_i corresponds to an x_i value or any y_i value. The derivatives, evaluated at $\underset{\sim}{U}$, are presented in Table 3.2 where

$$\begin{aligned} a &= \Sigma_{11}X_i - \Sigma_{12}X_i & e &= \Sigma_{11}X_i + \Sigma_{12}X_i \\ b &= \Sigma_{11}Y_i - \Sigma_{12}Y_i & f &= \Sigma_{21}X_i + \Sigma_{22}X_i \\ c &= \Sigma_{21}X_i - \Sigma_{22}X_i & g &= \Sigma_{11}Y_i + \Sigma_{12}Y_i \\ d &= \Sigma_{21}Y_i - \Sigma_{22}Y_i & h &= \Sigma_{21}Y_i + \Sigma_{22}Y_i \end{aligned} \tag{3.15}$$

Now substituting into equation (3.7) the Taylor series expansion for each function becomes

$$f(\underset{\sim}{u}) = f(\underset{\sim}{U}) + \sum_{i=1}^{N} \sum_{i=1}^{2} \left(\frac{\partial f(u)}{\partial u_i} \bigg|_{\underset{\sim}{U}} \right) (u_{ij} - U_{ij}) \tag{3.16}$$

and the variance given by equation (3.9) is

Table 3.2 Derivatives of $f_j(U)$ with Respect to U_i by Group, G_{ij}

$\underset{\sim}{U}$		$f_1(U)$	$f_2(U)$	$f_3(U)$	$f_4(U)$	$\partial f_5(U)$
$y_i \in$	G_{11}	$1/a$	$1/N_1 - \bar{X}_1/a$	0	0	$(A/g-B/b)/E-F) + F(A-B-C+d)/[b(E-F)^2]$
	G_{12}	$-1/a$	$1/N_1 + \bar{X}_1/a$	0	0	$(A/g+B/b)/(E-F) - F(A-B-C+d)/[b(E-F)^2]$
	G_{21}	0	0	$1/c$	$1/N_2 - \bar{X}_2/c$	$(-C/h+D/d)/(E-F) - E(A-B-C+D)/[d(E-F)^2]$
	G_{22}	0	0	$-1/c$	$1/N_2 + \bar{X}_2/c$	$(-C/h-D/d)/(E-F) + E(A-B-C+D)/[d(E-F)^2]$
$X_i \in$	G_{11}	$-b/a^2$	$-\beta_{11}/N_1 + \bar{X}_1 b/a^2$	0	0	$(A/A-B/e-C/a+D/a)/(E-F) - E(A-B-C+D)/[a(E-F)^2]$
	G_{12}	b/a^2	$-\beta_{11}/N_1 - \bar{X}_1 b/a^2$	0	0	$(-A/a-B/e+C/a-D/a)/(E-F) + E(A-B-C+D)/[a(E-F)^2]$
	G_{21}	0	0	$-d/c^2$	$-\beta_{21}/N_2 + \bar{X}_2 d/c^2$	$(A/c-B/c-C/c+D/f)/(E-F) + F(A-B-C+D)/[c(E-F)^2]$
	G_{22}	0	0	d/c^2	$-\beta_{21}/N_2 - \bar{X}_2 d/c^2$	$(-A/c+B/c+C/c+D/f)/(E-F) - F(A-B-C+D/[c(E-F)^2]$

$$\mathrm{Var}[f(\underset{\sim}{u})] = E\left\{\left[\sum_{i=1}^{N}\sum_{j=1}^{2}\left(\frac{\partial f(\underset{\sim}{u})}{\partial u_i}\bigg|_{\underset{\sim}{U}}\right)(u_{ij} - U_{ij})\right]^2\right\} . \tag{3.17}$$

Following the same reasoning which led from equation (3.9) to expression (3.12) gives

$$Z = \sum_{i=1}^{N}\left[\left(\frac{\partial f(\underset{\sim}{u})}{\partial u_i}\bigg|_{\underset{\sim}{U}}\right) x_i + \left(\frac{\partial f(\underset{\sim}{u})}{\partial u_i}\bigg|_{\underset{\sim}{U}}\right) y_i\right] \tag{3.18}$$

hence,

$$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{Var}\left\{\sum_{i=1}^{N}\left[\left(\frac{\partial f(\underset{\sim}{u})}{\partial u_i}\bigg|_{\underset{\sim}{U}}\right) x_i + \left(\frac{\partial f(\underset{\sim}{u})}{\partial u_i}\bigg|_{\underset{\sim}{U}}\right) y_i\right]\right\} \\ &= \mathrm{Var}\left\{\sum_{i=1}^{N} z_i\right\} .\end{aligned} \tag{3.19}$$

3.5.3 Pseudo Replication Technique

The problem now is to find an estimate of Var(Z). This would be a simple matter if there were replications of the data since Z could be calculated for each replicate. The variance could then be calculated as the variation between replicates.

Since there are no replicates for the model being studied, the only feasible approach is to borrow a technique from the field of sample survey methodology known as the collapsed strata method (Cochran, 1977; Hanson, Hurwitz and Madow, 1953; Hartley, Rao and Kiefer, 1969). This approach imposes a pseudo replication on the data.

In order to use this method assume that each observation (x_i,y_i) comes from a separate stratum. Hence, there are i=1,N strata each containing one observation. Next, group these strata into G groups

Lg strata in the g-th group. In this case, let the G groups be made up of adjacent observations all of which are in the same segment based on the prior knowledge about N_1. In addition, the observations in a group are either all below or all above the median point of the segment. The strata in a group can be viewed as Lg replications of the g-th group. For this model, Lg, the number of elements in a group, is usually two. Exceptions are those groups near the join point and the median points of the segments.

Now let t_{gh} be the z_i value as given in equation (3.12) for the h-th element in the g-th group. Note that

$$\sum_{i=1}^{N} z_i = \sum_{g=1}^{G} \sum_{h=1}^{L_g} t_{gh} \quad .$$

The variance of Z can then be estimated as (Cochran, 1977, p. 138)

$$\mathrm{Var}(Z) = \mathrm{Var}\left[\sum_g^G \sum_h^{L_g} t_{gh}\right] = \sum_g^G \frac{L_g}{L_g - 1} \sum_h^{L_g} (t_{gh} - \bar{t}_g)^2$$

where $\bar{t}_g = \sum_h^{L_g} t_{gh}/L_g$.

This estimate has a positive bias which is negligible if there is little difference in the true values of the strata in a group. Clearly, this is not the ideal approach to estimating a variance; however, it is the only feasible approach when there are no replicates and no information about the error variances.

3.6 SIMULATION PROCEDURE FOR THE EMPIRICAL DISTRIBUTION OF THE JOIN POINT

For large samples, the join point estimate was shown to be consistent (section 3.4) and a variance estimator was proposed (section 3.5). In order to study the behavior of these estimators for small samples a simulation study is needed to obtain the empirical distribution of the join point estimate and the asymptotic variance. In addition, the simulation results from this chapter will be used in chapter 4 to compare them with parallel results from the Hudson technique of chapter 2.

The model simulation is outlined in section 3.6.1. Next, the simulation of the variance estimation procedure is described in section 3.6.2. Finally, the method of calculating the empirical moments and the asymptotic variance is given in section 3.6.3.

3.6.1 Simulation of the Model

The same procedure described in chapter 2 was used to simulate 1000 samples of size $N = 60$ for use in studying the empirical distribution of the Wald-type estimators. For the case where N_1 is assumed to be known the Wald-type estimators were calculated and the derivatives needed for the variance estimate were based on these estimators of the regression parameters.

When there is no prior knowledge of the location of the join point an iterative procedure is simulated. The iterative procedure is parallel to that proposed by Hudson and described in section 2.4.1b except that Wald-type estimators were used for the regression coefficients rather than the ordinary least squares estimators.

3.6.2 Simulation of Variance Estimates for Each Sample

For each of the 1000 simulated samples the derivatives given in Table 3.2 were calculated. The parameter estimates obtained from the sample were substituted for the true values in the derivatives. Next, the z_i were calculated for each observed pair (x_i,y_i), $i = 1,N$. These z_i values were then allocated to groups as described in section 3.5.3 in order to obtain the t_{gh} values. The estimated variances were then calculated for each sample using equation (3.19).

3.6.3 Calculation of the Empirical Moments and Asymptotic Variance

The empirical moments were calculated by taking the mean and variance of the parameter estimates over all one thousand samples. Similarly, the asymptotic variance was found by averaging the variance estimates from each sample over all the samples.

3.7 RESULTS

As in chapter 2 the procedures discussed in this chapter are illustrated using a numerical model. The model and parameter values which are used are outlined in section 3.7.1. Section 3.7.2 presents the simulation results for the studies in which N_1 is assumed to be known. Next, the numerical variance estimates based on the simulation study are discussed in section 3.7.3. A comparison is made between the estimated variances and the empirical variances in section 3.7.4. Finally, the results of the simulated iterative procedure are presented in section 3.7.5.

3.7.1 Numerical Model and Parameter Values

The mathematical model under study is the same as that described in section 2.5, case 2. The following table summarizes the combinations of error variances and prior knowledge which will be studied for the Wald estimators.

Table 3.3 Parameter Values for Numerical Results

Parameters	Parameter Values
N	60
σ_ε^2	0.01, 0.1, 0.5, 1.0, 5.0
σ_ν^2	0.0, 0.01, 0.05, 0.10, 0.50, 1.0, 5.0
prior knowledge of X_{N_1}	13.0, 13.5144, none

While these combinations do not cover all the possibilities explored for the Hudson technique in the previous chapter they represent the most important cases needed for comparing the Hudson technique with the Wald technique.

Based on the results of this simulation several points about the empirical distribution of $\hat{\alpha}$ should be observable. First, the mean join point estimate over all 1000 samples should be close to the true join point value, 13.5144, for all values of σ_ν^2 and σ_ε^2 since the Wald-type estimators are consistent estimators. In addition, the variance of the sample should be close to the average asymptotic variance based on the Taylor series expansion as described in section 3.5.

3.7.2 Empirical Distribution from Simulation for N_1 Known

The simulation results for samples of size N = 60 and prior knowledge that X_{N_1} = 13.5144 are given in Table 3.4. Note first the mean join point estimate over the 1000 samples (col. 1). This mean is within 3 standard errors of the true join point for all values of σ_ε^2 and σ_ν^2. Hence, the join point estimate does appear to be asymptotically unbiased.

3.7.3 Numerical Variance Estimates from the Simulation Study

The asymptotic variances are based on numerical techniques described in section 3.5. The average of the estimates of the asymptotic variance for $\hat{\alpha}$ are presented in the last column (4) of Table 3.4. These variance estimates are the mean asymptotic variance over 1000 samples for various values of σ_ε^2 and σ_ν^2. As expected, the asymptotic variance increases as σ_ν^2 increases and also as σ_ε^2 increases.

The variance of the computed join points and the averaging of the estimated variances are close when σ_ν^2 is within the range which satisfies assumption (viii). Recall that the assumption states that ν_i should rarely exceed half the minimum distance between two data points. This assumption is violated when $\sigma_\nu^2 > 0.01$.

3.7.4 Improper Specification of N_1

In addition to the effects of increasing error, it is interesting to note the effect of improperly specifying the location of the join point. Table 3.5 presents results when X_{N_1} is specified to be 13 as well as results when X_{N_1} is 13.5144 and when X_{N_1} is unknown. When X_{N_1} is thought to be 13 and there is measurement error ($\sigma_\nu^2 > 0.0$) it appears that none of the mean estimates are within 3 standard errors of the true value and there are large differences between the asymptotic variances and the variances over the 1000 samples. By comparison, when the estimation relies on the iterative procedure all the estimates are within 3 standard errors of the true value.

3.7.5 Empirical Distribution for N_1 Unknown

The results of using an iterative procedure in order to obtain join point estimate when N_1 is not known are presented in the last part of Table 3.5. In addition to obtaining estimates which are within 3 standard errors of the true value for all σ_ν^2, the empirical variances for the iterative procedure are very close to the empirical variances when prior knowledge fixed x_{N_1} to be the true join point value. Unfortunately, the asymptotic theory is inapplicable when N_1 is unknown so it is impossible to obtain a variance estimate from a single sample.

3.8 CONCLUSIONS

It appears from the results persented in section 3.7 that the simulation procedure behaves as expected when the assumptions of the model are met. The mean join point estimate is within three standard errors of the true join point value, 13.5144. When the measurement error is small enough to insure that assumption (viii) is met, the variances of the computed join points are close to the mean of the asymptotic variances.

It would seem that an iterative approach is preferred for this estimation technique rather than taking the risk of misspecifying the value of X_{N_1}. A similar result was seen in chapter 2 for the Hudson technique. A comparison of the two techniques will be given in the next chapter.

Table 3.4

Statistics Describing the Distribution of the Join Point Estimates Over 1000 Samples of Size N=60 for Various Values of σ_ε^2 and σ_ν^2 When Prior Knowledge Specifies x_{N_1}=13.5144 and Estimates Are Based on the Wald Technique

σ_ε^2	σ_ν^2	Mean Over All Samples (1)	Standard Error (2)	Variance of Computed Join Points (3)	Mean of Asymptotic Variances (4)
0.01	0.00	13.5143	.000515	.000265	.000264
	0.01	13.5123	.001076	.001157	.001054
	0.05	13.5151	.002142	.004586	.021794
	0.10	13.5127	.003030	.009183	.126114
	0.50	13.5017	.006102	.037240	.499246
	1.00	13.5122	.008116	.065868	.702395
	5.00	13.5385	.015254	.232680	1.326890
0.10	0.00	13.5140	.001629	.002653	.002636
	0.01	13.5124	.001852	.003430	.003443
	0.05	13.5139	.002634	.006936	.024003
	0.10	13.5117	.003464	.012000	.131777
	0.50	13.5042	.006259	.039177	.503938
	1.00	13.5118	.008256	.068156	.703756
	5.00	13.5402	.015323	.234806	1.332900
0.50	0.00	13.5135	.003642	.013261	.013180
	0.01	13.5125	.003715	.013802	.014037
	0.05	13.5118	.004096	.016779	.034106
	0.10	13.5100	.004799	.023029	.146716
	0.50	13.5087	.006828	.046622	.514397
	1.00	13.5112	.008737	.076343	.716403
	5.00	13.5436	.015557	.242010	1.353570
1.00	0.00	13.5132	.005146	.026483	.026363
	0.01	13.5126	.005165	.026676	.027294
	0.05	13.5101	.005317	.028270	.046845
	0.10	13.5089	.005948	.035375	.162706
	0.50	13.5117	.007452	.055535	.533643
	1.00	13.5100	.009230	.085189	.737777
	5.00	13.5453	.015806	.249823	1.375780
5.00	0.00	13.5106	.009737	.094817	.131705
	0.01	13.5100	.009814	.096315	.133063
	0.05	13.4998	.009465	.089590	.151856
	0.10	13.5027	.010252	.105107	.278651
	0.50	13.5236	.010739	.115331	.678819
	1.00	13.5032	.011916	.141997	.884812
	5.00	13.5458	.017410	.303104	1.565090

Table 3.5

Statistics Describing the Distribution of the Join Point Estimate for Various Values of σ_ν^2 and the Prior Knowledge About N_1 when $\sigma_\varepsilon^2 = 0.01$ and Estimates Are Based on the Wald Technique

x_{N_1}	σ_ν^2	Mean Over All Samples (1)	Standard Error (2)	Variance of Computed Join Points (3)	Mean of Asymptotic Variances (4)
13.0000	0.00	13.5143	.000515	.000265	.000264
	0.01	13.2803	.006542	.042791	1.351890
	0.05	13.3248	.005848	.034197	1.351110
	0.10	13.3391	.005639	.031797	1.332690
	0.50	13.3358	.006711	.045044	1.102140
	1.00	13.3543	.008648	.074794	1.014290
	5.00	13.2943	.014970	.224096	1.478520
13.5144	0.00	13.5143	.000515	.000265	.000264
	0.01	13.5123	.001076	.001157	.001054
	0.05	13.5151	.002142	.004586	.021794
	0.10	13.5127	.003030	.009183	.126114
	0.50	13.5017	.006102	.037240	.499246
	1.00	13.5122	.008116	.065868	.702395
	5.00	13.5385	.015254	.232680	1.326890
none	0.00	13.5143	.000515	.000265	*
	0.01	13.5123	.001076	.001157	
	0.05	13.5153	.002180	.004753	
	0.10	13.5143	.003158	.009973	
	0.50	13.5068	.006871	.047216	
	1.00	13.5904	.047409	2.247590	
	5.00	13.6169	.070602	4.984610	

*not applicable

CHAPTER 4

A COMPARISON OF THE HUDSON AND WALD TECHNIQUES FOR JOIN POINT ESTIMATION IN PWLR MODELS

4.1 INTRODUCTION

The technique proposed by Hudson for estimating the join point in PWLR models was studied in chapter 2. The use of this technique was explored in the presence of measurement error in which case the OLS estimate is known to be biased. In chapter 3 the Wald grouping technique, which can be used to obtain consistent regression coefficient estimates in the presence of measurement error, was proposed as an alternative to the OLS technique for estimating the join point.

In this chapter a comparison is drawn in section 4.2 between the results from the Hudson and Wald techniques when applied to the same model. Next, in section 4.3, conclusions are drawn based on what type of knowledge one has concerning the data being analyzed. Finally, recommendations are made in section 4.4 concerning the use of the two estimation techniques.

4.2 COMPARISIONS OF THE WALD AND HUDSON TECHNIQUES

4.2.1 <u>Presentation of Results</u>

The comparable results from tables 2.5 and 3.4 are presented in Table 4.1. This table gives the means and variances over the thousand samples for the join point estimates. In addition, to aid in interpreting these results, the mean squared error (MSE) is presented in column (3). The MSE is calculated as the variance plus the square of the bias which is in turn the difference between the estimate and the true value.

4.2.2 Discussion of Compared Results

Look first at the case when $\sigma_\nu^2 = 0.0$. It can be seen that both procedures give unbiased estimates. However, the Hudson technique has perhaps a slightly lower variance and MSE than the Wald technique.

Next, turn to the case where x_{N_1} is specified to be 13.0 . Recall that if there is no measurement error then this value is the true value since the last abscissa before the join is 13.0 . However, as the measurement error increases, the ordering of the observed values may differ from the true value. Thus specifying $x_{N_1} = 13$ may lead to erroneous results due to misspecification of the observed values relative to the true values. In comparing, the MSE's for the case when $\sigma_\nu^2 > 0$ and $x_{N_1} = 13$ the Wald technique has slightly smaller MSE's in every case.

Now observe the results for the case when x_{N_1} is specified to be the true join point value, 13.5144 . This specification decreases the chance that the observed values will be classified in the wrong segment; however, it does not entirely eliminate the possibility. The MSE's for both techniques are smaller than those found for previous specification of $x_{N_1} = 13$. Between the Wald and Hudson techniques the MSE's are not consistently smaller for either technique.

Resulting MSE's for the iterative procedure are very similar to those for the case when $x_{N_1} = 13.5144$. In addition, the MSE's for both the Wald and Hudson techniques are very similar.

4.3 CONCLUSIONS

In forming conclusions about which technique is preferable in a particular analysis there are several points which need to be considered. Among these considerations are the knowledge one has about the true location of the join point and the extent to which there is measurement error in the independent variable.

4.3.1 Knowledge Concerning N_1

First, one must evaluate the knowledge available about the true location of the join point in order to determine whether or not an iterative procedure is needed. It seems clear for the model studied here that some type of iterative procedure is in order. It appears that using outside knowledge about the location of the join to avoid using iterative procedures is not the best approach at least for the model under study.

4.3.2 Knowledge Concerning Measurement Error

A second consideration is the available knowledge about the presence of measurement error in the data. In the simulation study the measurement error variance was allowed to range from 0 to 5.0 . In a realistic experiment or study it seems unlikely that data which has an error variance even as large as 1.0 would be considered when the intervals between the points of observation are on the order of 1.0 . Hence, in deciding between the Hudson and Wald techniques it does not seem unreasonable to restrict the discussion to cases where the measurement error is thought to be small enough so that all the assumptions hold for both methods. If this were the case it would seem that the Wald technique is preferable if there is reason to believe that there might be some measurement error in the data. The performance of the two techniques is similar based on the MSE with respect to the join point estimation. However, the Wald technique has the advantage of also providing consistent estimates of the regression coefficients as well.

4.4 RECOMMENDATIONS

In situations where it is thought that a model similar to the model given by equation (2.1) applies and there is reason to believe that measurement error exists in the independent variable then the Wald method should be used to estimate the join point. An iterative procedure gives the better results but can be used in conjunctin with prior knowledge about the location of the join point. Such prior knowledge about the approximate location of the join can be used to choose relevant values of x_{N_1} so that the iterative algorithm would not be required to process every possible value in the scope of the data.

Table 4.1 Simulation Results from Both the Hudson and Wald Techniques for Various Values of σ_ν^2 and Prior Knowledge About N_1 When $\sigma_\varepsilon^2 = 0.01$

		HUDSON			WALD		
x_{N_1}	σ_ν^2	Mean Over All Samples (1)	Variance Over All Samples (2)	MSE[1] (3)	Mean Over All Samples (4)	Variance Over All Samples (5)	MSE[1] (6)
13.0	0.00	13.5143	.000225	.0002	13.5143	.000265	.0003
	0.01	12.7833	.569584	1.1041	13.3500	.018614	.0456
	0.05	12.7831	.584922	1.1197	13.3798	.016434	.0346
	0.10	12.8278	.581208	1.0526	13.3886	.017041	.0329
	0.50	12.9227	.576459	0.9266	13.3820	.037141	.0547
	1.00	12.9104	.024736	.3896	13.3938	.071093	.0856
	5.00	12.9017	.027340	.4027	13.3622	.253564	.2767
13.5144	0.00	13.5143	.000225	.0002	13.5143	.000265	.0003
	0.01	13.5129	.000974	.0010	13.5123	.001157	.0012
	0.05	13.5098	.009140	.0092	13.5151	.004606	.0046
	0.10	13.5046	.012965	.0131	13.5130	.009304	.0093
	0.50	13.4623	.086994	.0897	13.5032	.038880	.0390
	1.00	13.4465	.011575	.0162	13.5126	.070928	.0709
	5.00	13.3841	.018520	.0355	13.5451	.272589	.2735
unknown	0.00	13.5143	.000225	.0002	13.5143	.000265	.0003
	0.01	13.5129	.000974	.0010	13.5123	.001157	.0012
	0.05	13.5144	.003908	.0039	13.5153	.004753	.0048
	0.10	13.5125	.007836	.0078	13.5143	.009973	.0100
	0.50	13.5232	.038628	.0387	13.5422	1.014760	1.0155
	1.00	13.5403	.009069	.0097	13.5892	2.247670	2.2533
	5.00	13.6164	.037265	.0477	13.6244	4.991390	5.0035

[1] MSE = (mean - 13.5144)^2 + variance

CHAPTER 5

PIECEWISE LINEAR REGRESSION WITH REPLICATED DATA

5.1 INTRODUCTION

When the data contain n_i observations x_{ij} on each of the N X_i values it is a simple matter to estimate the measurement error variance, σ_ν^2, as well as σ_ϵ^2 . This additional information can then be used in order to obtain consistent regression parameter estimates in the usual regression model in spite of errors in measurement.

One use of this additional information for the simple linear regression model was proposed by Tukey (1951) and clarified by Madansky (1959). Their works give several slope estimators which can be obtained using the partitioned sums of squares for the data. Slope estimators and the resulting join point estimators will be discussed in section 5.2.

Asymptotic variances of the slope estimators can be found using Taylor series expansions. In fact, the variance estimates are given in Dorff and Gurland (1961a). However, it is again difficult to obtain the variance of the join point estimate by means of Taylor series approximations. Obtaining a variance estimate for replicated data is beyond the scope of this dissertation and will not be addressed.

A simulation study is described in section 5.3 and the results of a numerical model are shown in section 5.4.

5.2 MODEL AND ESTIMATION PROCEDURE

5.2.1 Model under Study

The model with replicated data can be written:

$$y_{ij} = \begin{cases} \beta_{10} + \beta_{11}x_{ij} + \varepsilon_{ij} - \nu_{ij}\beta_{11} \; , \; X_i \leq \alpha \\ \\ \beta_{20} + \beta_{21}x_{ij} + \varepsilon_{ij} - \nu_{ij}\beta_{21} \; , \; X_i > \alpha \end{cases} \tag{5.1}$$

where there are $j = 1,2,\ldots,n_i$ replications for each of the $i = 1,\ldots,N$ true X_i values. Note that this is virtually the same model as that given by equation (2.1) with the addition of the index j for replication.

The assumptions of model (5.1) remain the same as those for model (2.1) and are given in section 2.2.1 . However, it is necessary to say something more about the independence of the ν_{ij} and the ε_{ij} when there are replications. In the case of replicated data assume that ν_{ih} is independent of ν_{jk} unless $i = j$ and $h = k$ and that ε_{ih} is independent of ν_{jk} for all i,j,k and h.

Assumption (vii) which was introduced in the previous chapter (section 3.2.1) is not needed for this model and the estimation procedures which are used. However, assumption (viii) discussed in section (3.3.2) must be retained to insure that the ordering of the x_i is the same as that of the X_i. Note, however, that assumption (viii) is no longer needed to insure the consistency of the slope estimates -- only to insure that N_1 is properly specified.

5.2.2 Estimation Procedure

The easiest way to present the slope and variance estimators based on replicated data is to begin by looking at the ANOVA table -- Table

Table 5.1

ANOVA Table for a Piecewise Regression
Model with k = 2 Segments

Source		
$B_{xx(k)}$	$\sum_{i=1}^{N_k} n_i(\bar{x}_{i\cdot}-\bar{x}_{(k)})^2/(N_k-1)$	$\sigma_\nu^2 + [(n^2-\sum_{i=1}^{N_k} n_i^2)/N_k n_k-n_k] s_X^2(k)$
$B_{xy(k)}$	$\sum_{i=1}^{N_k} (\bar{x}_{i\cdot}-\bar{x}_{(k)})(\bar{y}_i-\bar{y}_{(k)})/(N_k-1)$	$[(n_k^2-\sum_{i=1}^{N_k} n_i^2)/(N_k n_k-n_k)]\beta_{k1} s^2_{X(k)}$
$B_{yy(k)}$	$\sum_{i=1}^{N_k} n_i(\bar{y}_{i\cdot}-\bar{y}_{(k)})^2/(N_k-1)$	$\sigma_\varepsilon^2+[(n^2-\sum_{i=1}^{N_k} n_i^2)/(N_k n_k-n_k)]\beta_{k1}^2 s^2_{X(k)}$
$W_{xx(k)}$	$\sum_{i=1}^{N_k}\sum_{i=1}^{n_i} (x_{ij}-\bar{x}_{i\cdot})^2/(n_k-N_k)$	σ_ν^2
$W_{yy(k)}$	$\sum_{i=1}^{N_k}\sum_{j=1}^{n_i} (y_{ij}-\bar{y}_{i\cdot})^2/(n_k-N_k)$	σ_ε^2
W_{xx}	$\sum_{i=1}^{N}\sum_{j=1}^{n_i} (x_{ij}-\bar{x}_{i\cdot})^2/(n-N)$	σ_ν^2
W_{yy}	$\sum_{i=1}^{N}\sum_{j=1}^{n_i} (y_{ij}-\bar{y}_{i\cdot})^2/(n-N)$	σ_ε^2

where

$$n_k = \sum_{i=1}^{N_k} n_i \qquad n = \sum_{k=1}^{2} n_k$$

$$\bar{x}_{i\cdot} = \sum_{j=1}^{N_k} x_{ij}/n_i \qquad \bar{y}_{i\cdot} = \sum_{j=1}^{N_i} y_{ij}/n_i$$

$$\bar{x}_{(k)} = \sum_{i=1}^{N_k}\sum_{j=1}^{N_i} x_{ij}/n_k \qquad \bar{y}_{(k)} = \sum_{i=1}^{N_k}\sum_{j=1}^{N_i} y_{ij}/n_k$$

5.1 . Note that, in addition to the between and within sums of squares for each segment, the within sums of squares are calculated from the entire sample as well.

The estimation of the error variances will be briefly presented first in section 5.2.2a. Then, the estimation of the regression coefficients and the join point will be discussed in section 5.2.2b.

5.2.2a Estimation of Error Variances. When the data are replicated it is possible to directly estimate the error variances from the within sums of squares. Notice that in the ANOVA table the expected value of the mean square within the observed x for the k^{th} segment is σ_ε^2; and the expected value of the mean square within the observed y for the k^{th} segment is σ_ν^2 . Thus these mean squares provide unbiased estimates of the error variances. Note that in addition to the between and within mean squares for each of the k = 1,2 segments, the within mean squares are calculated from the entire sample as well. With large enough sample sizes the estimates of σ_ν^2 and σ_ε^2 should be the same no matter which within mean square is used in the estimation. For the purpose of this study W_{xx} and W_{yy} will be used in all estimates rather than $W_{xx}(k)$ and $W_{yy}(k)$ since they are based on more information and, therefore, likely to be more accurate.

5.2.2b Estimation of Regression Parameters. Tukey (1951) proposed three different consistent slope estimators based on the mean squares seen in Table 5.1. It can be seen that all of the following estimators converge to β_{k1} as the number of replications and N_1 and N_2 go to infinity.

$$ {}_1b_{k1} = B_{xy}(k)/[B_{xx}(k) - W_{xx}] \tag{5.2} $$

$$ {}_2b_{k1} = [B_{yy}(k) - W_{yy}]/B_{xy}(k) \tag{5.3} $$

$$_3b_{k1} = \{[B_{yy}(k) - W_{yy}]/[B_{xx}(k) - W_{xx}]\}^{\frac{1}{2}} \tag{5.4}$$

$$\text{where } \operatorname{sgn}(_3b_{k1}) = \operatorname{sgn}[B_{xy}(k)] \ .$$

Both β_{k0} and α are estimated in the usual way by:

$$b_{k0} = \bar{y}_{(k)} - b_{k1} \cdot \bar{x}_{(k)} \qquad \text{and} \tag{5.5}$$

$$a = (b_{10} - b_{20})/(b_{21} - b_{11}) \tag{5.6}$$

where b_{k1} can be substituted from any of the estimates given in equations (5.2), (5.3) and (5.4).

The variance of 'a' can be obtained using the Tepping approach to the Taylorized sum. Since the data are replicated in this case, it would not be necessary to use the pseudo replication method introduced in chapter 2. However, the question of a variance estimate will not be addressed in this study.

5.3 SIMULATION PROCEDURE

In order to study the empirical distribution of the Tukey estimators and compare them with the Hudson and Wald estimators a simulation study was done. The simulation procedure is similar to that described in section 2.4 . The main difference between the simulation of the model in section 2.4 and this case is that for each of the N data points n_i replications were also generated to produce a total of $n_i N$ observations.

For the case where N_1 is assumed to be known it is necessary to divide all $n_i N$ observations into the two segments. In order to accomplish this division into segments, the data were sorted on the $\bar{x}_{i.}$ values so that all those data points corresponding to $\bar{x}_{i.} \leq x_{N_1}$ formed the first segment while the remaining points formed the second segment. This approach was taken because $\bar{x}_{i.}$ is an unbiased estimate of X_i.

5.4 RESULTS

5.4.1 Numerical Model

The mathematical model being studied is again the same as the one described in section 2.5 . For each of the N = 60 data points, $n_i = 6$ replications were generated to give a total of 360 observations. Each of the three different slope estimates were used to obtain a join point estimate for the following parameters:

Table 5.2 Parameter Values for Numeric Results

Parameter	Values
σ_ε^2	0.01, 0.10, 0.50
σ_ν^2	0.00, 0.01, 0.05, 0.10, 0.50
prior knowledge of x_{N_1}	13.0, 13.5144

Note that the error variances studied here do not get as large as those studied in chapters 2 and 3. It was seen in chapter 3 that $\sigma_\nu^2 > 0.01$ actually violated assumption (viii). Therefore, it was felt that results for variances as large as 1.0 and 5.0 were unnecessary.

5.4.2 Estimation of the Error Variances

It was seen in section 5.2.2a that estimates of the error variances could be easily estimated from the within mean squares. The average over all samples of these estimates for the simulation study are given in Table 5.3 along with the true error variances specified in the simulation. The fact that the estimates are not precise is a function of the GAUSS subroutine used to generate the error terms. As an experiment several runs were made with more replications in order to see if the estimate is more precise with a larger sample size. It can be seen at the bottom of Table 5.3 that increasing the number of replications does improve the estimate for σ_ε^2 slightly, but it still has a negative bias.

5.4.3 Estimation of the Join Point when N_1 Is Known

The simulation results for all three estimators and all combinations of different error variances are given in Table 5.4 for the prior knowledge specified to be 13.5144 . Note first that there are very few differences between the means and variances for the three different estimators. Also when differences do occur between the estimates they are only very small ones.

Looking now at the means of the computed coefficients, it can be seen that the bias is very small even for large error variances. Recall that this was not the case for either the Hudson technique which was known to be biased nor for the Wald technique which was shown to be asymptotically consistent. The reason, of course, is that these estimates are based on much more information since the total sample size is larger and the observations are replicated.

On the other hand, the variances over all samples are a good deal larger for the Tukey estimators than variances seen with either the Hudson or Wald technique. This result is not unexpected and can be attributed, at least in part, to the fact that the estimated error variances are used in determining the join point estimates.

Turning now to Table 5.5, the effect of misspecifying the prior knowledge about the location can be seen. When measurement error is present the means of the join point estimates over all samples is slightly lower than the means seen in Table 5.4 . Thus, while the measurement error effect is accounted for in the estimation technique, the misspecification of N_1 leads to biased estimates.

Currently, there is no iterative technique available for replicated data, but further study and development of an iterative algorithm would be useful.

5.5 CONCLUSIONS

Replicated data provide a much more feasible way of estimating error variances and parameters from the regression model. However, it is not always possible to obtain replicates, hence the main thrust of this study has been centered on estimation techniques which do not require replications. Clearly, if replications are available, information provided by them should be used.

Table 5.3

Estimated Error Variances Based on Tukey
Technique of Slope Estimation

6 replications

σ_ε^2	$\hat{\sigma}_\varepsilon^2$	σ_ν^2	$\hat{\sigma}_\nu^2$
0.01	.00849	0.01	.00847
	.00849	0.05	.04239
	.00844	0.10	.08464
	.00846	0.50	.42384
0.05	.04244	0.01	.00847
	.04243	0.05	.04239
	.04222	0.10	.08464
	.04244	0.50	.42368
0.10	.08488	0.01	.00847
	.08486	0.05	.04239
	.08444	0.10	.08464
	.08486	0.50	.42389
0.50	.42439	0.01	.00847
	.42428	0.05	.04239
	.42218	0.10	.08464
	.42218	0.50	.42319

10 replications

.01	.00914

20 replications

.01	.00966

Table 5.4 Statistics Describing the Distribution of the Join Point Estimate Over 1000 Samples

		α_1		α_2		α_3	
σ^2_ε	σ^2_ν	Mean of Computed Coefficient	Variance of Computed Coefficient	Mean of Computed Coefficient	Variance of Computed Coefficient	Mean of Computed Coefficient	Variance of Computed Coefficient
0.01	0.00	13.5146	.182863	13.5146	.182863	13.5146	.182863
	0.01	13.5140	.182964	13.5140	.182963	13.5140	.182964
	0.05	13.5150	.183449	13.5150	.183449	13.5150	.183449
	0.10	13.5122	.183976	13.5122	.183976	13.5122	.183976
	0.50	13.5117	.189753	13.5116	.187948	13.5116	.187948
0.05	0.00	13.5147	.183012	13.5147	.183012	13.5147	.183012
	0.01	13.5141	.183112	13.5140	.183112	13.5140	.183112
	0.05	13.5150	.183582	13.5149	.183582	13.5149	.183582
	0.10	13.5119	.184119	13.5119	.184119	13.5119	.184119
	0.50	13.5129	.188666	13.5127	.188648	13.5128	.188654
0.10	0.00	13.5148	.183198	13.5148	.183198	13.5148	.183198
	0.01	13.5141	.183295	13.5140	.183296	13.5141	.183295
	0.05	13.5149	.183747	13.5149	.183746	13.5149	.183746
	0.10	13.5117	.184301	13.5117	.184301	13.5117	.184301
	0.50	13.5167	.188835	13.5165	.188840	13.5166	.188835
0.50	0.00	13.5152	.184672	13.5152	.184671	13.5152	.184671
	0.01	13.5141	.184751	13.5141	.184760	13.5141	.184755
	0.05	13.5148	.185057	13.5147	.185057	13.5148	.185056
	0.10	13.5109	.185774	13.5108	.185776	13.5108	.185774
	0.50	13.5096	.190361	13.5093	.190361	13.5095	.190357

Sample Size: $N = 60$
Replicates: $n_1 = 6$
Prior Knowledge: $X_{N_1} = 13.5144$
Technique: Tukey

Table 5.5 Statistics Describing the Distribution of the Join Point Estimate Over 1000 Samples

		α_1		α_2		α_3	
σ_ε^2	σ_ν^2	Mean of Computed Coefficient	Variance of Computed Coefficient	Mean of Computed Coefficient	Variance of Computed Coefficient	Mean of Computed Coefficient	Variance of Computed Coefficient
0.01	0.0	13.5146	.182863				
	0.01	13.4861	.182871	13.4871	.182846	13.4866	.182858
	0.05	13.4873	.183242	13.4882	.183224	13.4878	.183233
	0.10	13.4838	.183622	13.4846	.183612	13.4842	.183617
	0.50	13.4832	.187303	13.4836	.187295	13.4834	.187296
0.05	0.00	13.5147	.183012	13.5147	.183012	13.5147	.183012
	0.01	13.4859	.183022	13.4870	.182997	13.4865	.183009
	0.05	13.4873	.183384	13.4882	.183365	13.4878	.183374
	0.10	13.4836	.183777	13.4844	.183767	13.4840	.183772
	0.50	13.4820	.188019	13.4823	.187998	13.4821	.188005
0.10	0.00	13.5148	.183198	13.5148	.183198	13.5148	.183198
	0.01	13.4858	.183212	13.4869	.183187	13.4863	.183199
	0.05	13.4873	.183560	13.4882	.183540	13.4878	.183550
	0.10	13.4835	.183971	13.4843	.183962	13.4839	.183966
	0.50	13.4874	.188405	13.4877	.188410	13.4875	.188405
0.50	0.00	13.5152	.184672	13.5152	.184671	13.5152	.184671
	0.01	13.4853	.184737	13.4863	.184717	13.4858	.184726
	0.05	13.4873	.184958	13.4881	.184931	13.4877	.184944
	0.10	13.4897	.185529	13.4837	.185521	13.4833	.185524
	0.50	13.4796	.189804	13.4798	.189813	13.4797	.189804

Sample Size: $N = 60$
Replicates: $n_i = 6$
Prior Knowledge: $X_{N_1} = 13.00$
Technique: Tukey

CHAPTER 6

APPLICATION

6.1 INTRODUCTION

The original motivation behind this research came from a desire to study the timing of ovulation in women as it is predicted by the waking, or basal, body temperature (BBT). Thus as an illustration of the estimation techniques developed in chapters 2 and 3, these techniques are applied to BBT data in an attempt to estimate the data of phase shift as the join point of a PWLR model.

6.1.1 Historical Background

Sources (Döring, Marshall, Vollman) differ slightly on the history of knowledge and use of BBT. Squire in 1868 seems to be the first person to note the pattern of temperature shifts from a low phase to a high phase in the middle of the menstrual cycle. The first critical experiments on fluctuations during the menstrual cycle are attributed to Mary Jacobi in 1876. It was not until 1928 that the claim of Van de Velde that this shift in phase is associated with ovulation was substantiated. Later studies, for example Martin (1943), linked the phase shift to elevated levels of the hormone progesterone and other clinical indicators of ovulation. In 1947, Ferin suggested that knowledge of this shift in BBT could be used in planning pregnancies. A German priest, Hillebrand, is given credit for developing the most reliable form of birth control based on periodic abstinence using BBT. In 1968, Marshall conducted a prospective study of the use of BBT information as a contraceptive method by conducting a field trial of the method.

6.1.2 Physiological Process

A menstrual cycle is defined as the day on which menstrual flow begins through the last day before menstruation begins again. Most normally menstruating women follow a similar pattern of hormonal changes during this cycle. In the post-menstrual phase, the time immediately after menstruation ends, plasma estrogen concentrations are relatively low and constant. A peak occurs in the plasma estrogen levels one to three days prior to ovulation. This rise in estrogen is believed to trigger a mid-cycle peak in luterizing hormone (LH) followed closely by ovulation. Once ovulation occurs there is a sharp rise in plasma progesterone levels which is linked to increases in the basal body temperature.

6.1.3 Analysis of the BBT Shift

While all fertile women seem to follow the same sequence of events physiologically, detection of these events, especially the change in BBT, is not always straightforward because of variability between women and between cycles and the sensitivity of the BBT to events besides the rise in plasma progesterone such as illness.

At this point it is well established (Döring, 1973) that conception never occurs between the third day after the temperature rise until the following menstruation. Thus there are only about 6 to 10 days, depending on age, in the premenstrual phase which are certain to be infertile days (Vollman, 1977). Rules to determine infertile days during the post-menstrual phase are much less certain. Two methods currently in use to determine the post-menstrual infertile days rely on taking cycles over several months and deducting days from the shortest cycle or, alternatively, from the day of earliest temperature shift in order to establish "safe" or infertile days. Both these methods attempt to deal with the variability between cycles though neither method is extremely reliable.

A 1966 scientific meeting sponsored by WHO has set up criterion for defining where the BBT shift occurs. The definition states that a significant shift is one which occurs in 48 hours or less, and in which three consecutive daily temperatures are at least 0.2° C higher than the last six daily temperatures prior to the shift. Since the shift in BBT is not always so obvious, Royston and Abrams (1980) propose a cumulative sum (CUSUM) test as an objective way of establishing when the shift has occurred. More recently, Carter and Blight (1981) present a sequential Bayes solution to the problem of detecting changes in BBT when viewed as a change-point regression problem.

The above mentioned procedures are all used to detect the date of phase shift, and therefore, to some extent, ovulation, after it has occurred. An important step towards predicting infertile days in the post-menstrual phase as well as determining the shift would be a retrospective statistical analysis of a large mumber of menstrual cycles. Such an analysis of variability in cycles could be used in establishing more solid rules for determining infertile days throughout a cycle. It is with this idea in mind that the piecewise regression model is proposed for the ovulation data.

In section 6.2 a description of the data used for this application is given. Next, the model and assumptions are discussed in section 6.3. Section 6.4 presents results from applying the Hudson and Wald techniques to the data in order to obtain join point estimates. Finally in section 6.5 recommendations are made concerning the use of PWLR to estimate shifts in BBT.

6.2 DESCRIPTION OF THE DATA

As an example of the use of a PWLR model as a possible statistical model for BBT, data from the "Comparative Study of Two Methods of Natural Family Planning" were used. This study was conducted by the University of North Carolina at Chapel Hill and Cedars-Sinai Medical Center, Los Angeles, from 1976 to 1978. The objective of the original study was to determine the efficacy of the Ovulation Method and the Sympto-Thermal Method of birth control in preventing pregnancy. Women in the Sympto-Thermal part of the study were instructed to take their temperatures orally immediately upon waking each day and record the temperature as well as other pertinent information on standard charts provided by the project. These charts were then collected at a central location. Specific information about the study design can be obtained from Wade, et al. (1981).

For the purpose here of attempting to model the phase shift, five charts were selected from among those collected by the Comparative Study. The data on BBT from these charts appear in Table 6.1 . No attempt was made to choose these charts at random or by any other probability sampling scheme. Rather, these charts were specifically chosen because of the obvious well-defined shift. The definition of where the shift occurs is roughly based on criteria set forth by WHO. Not all cases chosen adhere strictly to the WHO definition, but all have clear shifts. The reason for selecting only charts with clear shifts is that it was felt that if this model proved inadequate for an obvious shift then it would require a reevaluation and modification before being useful for cycles which had less obvious shifts. This study is not intended to be definitive in terms of estimating the day on which the shift occurs. Rather, it is a first step in finding a model which does accomplish this goal while dealing with the measurement error problem.

6.3 STATISTICAL FRAMEWORK AND MODELING PROCEDURE

The statistical model under study is the two segment PWLR model set forth in equation 2.1 . The independent variable is the day of the cycle, where day 1 is the first day of menstruation. The dependent variable is the BBT and α is the day on which the shift occurs.

Under the assumptions discussed in sections 2.2 and 3.2 assume that the true relationship between basal body temperature (BBT) and day in cycle (DAY) is given by:

$$BBT = \begin{cases} \beta_{10} + \beta_{11}(DAY) & , \quad DAY \leq \text{shift} \\ \beta_{20} + \beta_{21}(DAY) & , \quad DAY \geq \text{shift} \end{cases}$$

Assume also that day in cycle is fixed, but that there is an error corresponding to fluctuations around a set time each day. This fluctuation has a mean 0 and an unknown variance σ_ν^2 .

At first glance it appears that the independent variable, day of cycle, is fixed and there should be no problem using ordinary estimation procedures. However, on further consideration, one realizes that, while the recorded day is fixed, the actual time of waking fluctuates. It is known that the basal body temperature fluctuates throughout the day, independent of wake or sleep. Hence, error is introduced into the problem if the temperature is not taken at the same hour each day. There is, of course, no way to check for such error since the exact time of day is not usually recorded. The fact that such errors very likely exist makes it necessary to attempt to account for them in the analysis. The use of the technique proposed in chapter 3 will do this.

A further assumption pertains to prior knowledge about the value of N_1. By examining each chart, it is possible to establish N_1 and thus satisfy the assumption that N_1 is known for each case. However, the purpose of this experiment is to model the BBT shift without the necessity of studying each chart. Hence, the iterative procedure described in chapters 2 and 3 will be applied. Furthermore, because N_1 is not fixed, it is impossible to obtain variance approximations for the parameter estimates.

6.4 RESULTS

Parameter estimates were first obtained for each of the five cycles using all data for each entire cycle. Since the BBT must drop from the high or premenstrual phase to the low or post-menstrual phase at some point during menstruation it seems reasonable to exclude the entire menstrual period from the study in order to avoid the possibility of high phase values in the low phase of the cycle. In addition a third search was done restricting the iterative algorithm from day 10 to day 20.

6.4.1 Hudson Technique

The parameter estimates obtained by fitting the PWLR model to the BBT data for the entire cycle and the reduced cycle using the Hudson technique are given in Table 6.2 . The columns of specific interest are (5) and (6) because these are the shift estimate from the model and the actual day the shift occurred as read directly from the graph.

It can be seen in Table 6.2 that none of the three estimates is very close to the true value. In most cases, restricting the cycle to days after menstruation and searching for a join only between days 10 and 20 improves the estimate somewhat.

6.4.2 Wald Technique

Results for the Wald technique are given in Table 6.3 . An additional restricted search was made from day 15 to day 25 in an attempt to obtain a better estimate of the day the BBT shift occurred.

Theoretically, if measurement error is present, the Wald estimates should be nearer the true day of the shift than the Hudson estimators since the latter are biased. For searches labeled A and B the Wald procedure does not appear to be consistently better than Hudson. However, when the search is restricted to relevant days (C) then the Wald technique does appear to perform better in the majority of cases.

Changing the restriction for the Wald technique from days 5 to 15 to days 10 to 20 helps the estimate in some cases, but not in others. It seems more study is needed in order to apply a general rule for restricting the iteration.

6.5 RECOMMENDATIONS

The purpose of this example was to illustrate a possible use for the segmented regression model which had been studied in modelling BBT during a menstrual cycle. It seems clear that this model with two straight lines constrained to meet at the join point does not, in fact, fit these data. However, this does not rule out the use of other segmented models to analyze these data. Vollman suggests that BBT measured rectally follows a sinuous curve and that the oral temperature do also to a lesser extent. Perhaps, instead of linear segments constrained to join, the segments should be modelled as separate curves which are constrained to meet. A second possibility is to keep the idea of linear segments and drop the constraint that the segments meet.

This latter type of segmented regression is similar to the work done by Quandt. While these suggestions offer possible solutions to the problem of estimating the phase shift and modeling BBT over a menstrual cycle implementing them is beyond the scope of this dissertation and will not be attempted here.

It is disappointing that the two segment PWLR model did not perform wonderfully for estimating the shift in BBT. However, there are other applications in which this type of segmented model does well. For instance, some dose-response type situations where it is desired to estimate threshold levels of toxins as in Hasselblad, Creason and Nelson (1976) require segmented regression as studied here and there is potential for measurement error to be a problem in such studies. A second example where PWLR is applicable and a potential for measurement error exists is in biological studies such as the work of Mellits (1968) in studying total body water as a function of weight and height.

Table 6.1

Data on Basal Body Temperatures
for Five Cycles

	Cycle				
Day	1	2	3	4	5
1	97.9	98.0	97.8	97.8	98.0
2	98.0	97.9	97.4	97.2	97.2
3	98.2	97.4	97.5	97.4	97.2
4	97.6	97.5	97.5	97.6	97.3
5	97.9	97.7	97.6	97.2	97.3
6	97.9[1]	97.8	97.7	97.4	97.3
7	97.4	97.8	97.4	97.0	97.2
8	97.8	97.8	97.4	97.8	97.4
9	97.5	97.8	97.5	97.4	97.4
10	97.6	97.7	97.8	97.6	97.5
11	97.9	97.8	97.5	97.6	97.4
12	97.3	97.8	97.3	97.6	97.4
13	97.3	97.9	97.3	97.8	97.5
14	97.6	97.7	97.5	97.0	97.5
15	97.8	97.7	98.0	97.4	97.7
16	97.7	97.2	97.9	97.4	97.7*
17	97.5	97.4	98.0*	98.0*	98.4
18	97.7	97.7	98.3	98.0	98.5
19	97.4	97.3	98.4	98.0	98.6
20	97.5	97.7	98.2	98.4	98.6
21	97.4	97.9	98.3	98.2	98.6
22	97.3	97.9*	98.2	98.2	98.6
23	97.6*	98.1	98.7	98.4	98.5
24	98.1	98.3	98.5	98.2	98.6
25	98.0	98.2	98.5	98.2	98.6
26	98.1	98.2	98.5	98.6	98.7
27	98.1	98.7		98.4	98.6
28	98.0	98.6		98.0	98.3
29	98.4	98.6		98.2	
30	98.1	98.6		98.8	
31	98.2				
32	98.3				
33	98.3				
34	98.2				
35	98.3				

*Shift from low phase to high phase.

[1]Line indicates end of menstruation.

Source: "Comparative Study of Two Methods of Natural Family Planning", 1978, UNC and Cedars-Sinai Medical Center, Los Angeles, California.

Table 6.2

Parameter Estimates Based on the Hudson Technique
Applied to BBT Data for Five Cycles

Cycle		$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{20}$	$\hat{\beta}_{21}$	$\hat{\alpha}$	True Shift	Sample Size
1	A[1]	97.95	-0.03	96.64	0.05	16.74	23-24	35
	B[2]	97.86	-0.03	96.91	0.04	13.04		29
	C[3]	97.86	-0.03	96.91	0.04	13.04		
2	A	98.37	-0.03	97.32	0.03	3.12	22-23	30
	B	98.19	-0.04	95.69	0.10	17.41		24
	C	98.19	-0.04	95.69	0.10	17.41		
3	A	97.63	-0.02	96.66	0.08	9.98	16-17	26
	B	97.87	-0.05	96.66	0.08	9.49		22
	C	97.63	-0.01	96.46	0.09	12.00		
4	A	97.61	-0.05	96.92	0.06	6.58	17-18	30
	B	96.90	0.06	97.84	0.02	25.54		25
	C	97.23	0.03	96.72	0.06	14.00		
5	A	98.80	-0.80	96.64	0.08	2.45	16-17	28
	B	96.53	0.09	99.21	-0.03	22.21		24
	C	96.49	0.10	98.23	0.01	20.00		

[1] A = iterative search over entire cycle

[2] B = iterative search over cycle minus menstrual period

[3] C = restricted search from day 10 to day 20 for cycle minus period

Table 6.3

Parameter Estimates Based on Wald Technique from PWLR Model
Applied to BBT Data for Five Cycles

Cycle		$\tilde{\beta}_{10}$	$\tilde{\beta}_{11}$	$\tilde{\beta}_{20}$	$\tilde{\beta}_{21}$	$\tilde{\alpha}$	True Shift	Sample Size
1	A[1]	97.91	-0.02	96.11	0.07	19.41	23-24	35
	B[2]	97.12	0.03	0.0	2.81	35.00		29
	C[3]	97.77	-0.02	95.99	0.07	19.00		
	D[4]	97.51	0.004	95.69	0.08	23.00		
2	A	97.79	-0.01	95.66	0.10	18.97	22-23	30
	B	98.09	-0.03	95.68	0.10	17.75		24
	C	98.09	-0.03	95.68	0.10	17.75		
	D	98.09	-0.03	95.68	0.10	17.75		
3	A	98.20	-0.40	96.97	0.06	2.65	16-17	26
	B	97.99	-0.07	96.55	0.08	9.56		22
	C	97.54	-0.003	96.10	0.10	14.00		
	D	96.83	0.07	98.50	0.00	23.11		
4	A	97.39	0.01	96.79	0.06	12.50	17-18	30
	B	96.79	0.06	98.40	0.00	25.19		25
	C	97.48	-0.01	96.03	0.09	16.00		
	D	96.79	0.06	98.40	0.00	25.19		
5	A	98.80	-0.80	96.64	0.08	2.45	16-17	28
	B	96.56	0.09	98.77	-0.01	21.80		24
	C	96.52	0.11	98.10	0.02	17.00		
	D	96.56	0.09	98.77	-0.01	21.80		

[1]A = iterative search over entire cycle

[2]B = iterative search over entire cycle minus menstrual period

[3]C = restricted search from day 10 to day 20 over cycle minus period

[4]D = restricted search from day 15 to day 25 over cycle minus period

CHAPTER 7

SUGGESTIONS FOR FURTHER RESEARCH

7.1 EXTENSIONS OF THE MODEL

The model studied throughout this thesis is the simplest PWLR model possible -- that of two linear segments constrained to join. One possible extension of this model is a model with more than two linear segments and, therefore, more than one join to be estimated. Hudson (1960), Bellman and Roth (1969), and Hawkins (1976) all discuss this more complicated model; however, a study of measurement error effects has not been carried out to date.

A second extension of the simple model would be to include more than one independent variable. The $E(y)$ then becomes a multidimensional response surface which can be divided into sections of half planes. Fuller (1969) fits polynomials to such a three dimensional model; his work can be extended to include the possibility of measurement error.

A further extension of the two segment model is to study a case where the segments are not linear, but are some more complicated function of the independent variable. Hudson (1966) introduces such models with Type III joins but does not treat them as thoroughly as the simpler models. Gallant and Fuller (1973) also look at the Type III join and propose an alternate estimation technique. Lerman (1980) also treates more complicated functions in proposing a grid search method for obtaining join point estimates. The effects of measurement error on any one of these techniques requires further study.

7.2 CHANGES IN THE ASSUMPTIONS

The assumptions made concerning the model under study are fairly basic to regression analysis. However, the assumption of uncorrelated error terms could be extended to include study of the effects of measurement error when the errors are correlated. A further extension of the assumptions is to impose a distributional assumption on the error terms. In this study the theoretical results made no such assumptions while the simulation study used normally distributed error terms. Other possibilities include the uniform or Poisson distribution for the error terms.

7.3 FURTHER WORK ON THE ESTIMATION PROCEDURE

Immediate extensions of the work done in this thesis apply to the case where data is replicated. A variance estimator needs to be pinned down for the join point estimate based on the Tukey type slope estimates. An approach which would yield this estimate is that proposed by Tepping and described in chapter 3. A further extension of the estimation procedure is to develop an iterative search procedure for the Tukey approach. Since knowledge about the location is not always available some criterion for determining the best fit when data is replicated would be useful.

7.4 EXTENSIONS OF THE SIMULATION PROCEDURE

Mention was made in chapter 1 that there are several variations of the grouping procedure originally proposed by Wald (1940) for finding consistent slope estimates. Specifically, Bartlett (1949) proposes using tertiles to divide the data into groups and Lindley (1947) proposes the use of quartiles. A second approach based on quartiles, was proposed by Nair and Banerjee (1943). A study of the effects of measurement error on these alternative grouping estimators would be useful in determining which is more effective in the piecewise situation.

REFERENCES

Bartlett, M.S.; 1949, "Fitting a Straight Line when Both Variables are Subject to Error", BIOMETRICS, p. 207-212.

Bellman, Richard and R. Roth; 1969, "Curve Fitting by Segmented Straight Lines", JASA 64:1079-1084.

Blalock, H.M., C.S. Wells, and L.F. Carter; 1970, "Statistical Estimation with Random Measurement Error", p. 75 in SOCIOLOGICAL METHODOLOGY, ed. Borgath, E.F. and G.W. Bohrnstedt, S.F., Jossey-Bass, Inc., 1970.

Carleson, F.D., E. Sobel, and G.S. Watson; 1966, "Linear Relationship Between Variables Affected by Error", BIOMETRICS 22:252-267.

Carter, L.F. and H.M. Blalock; 1970, "Underestimation of Error in Wald-Barlett Slope Estimation", JRSS Series C 19 (Part 1):34-41.

Carter, R.L. and B.J.N. Blight; 1981, "A Bayesian Change-Point Problem with an Application to the Prediction and Detection of Ovulation in Women", BIOMETRICS 37:743-751.

Cochran, W.G.; 1968, "Errors of Measurement in Statistics", TECHNOMETRICS 10(4):637-666.

———; 1977, SAMPLING TECHNIQUES, Wiley and Sons, Inc., New York.

Döring, Gerhard L.; 1973, "Detection of Ovulation by the Basal Body Temperature Method", PROCEEDINGS OF A RESEARCH CONFERENCE ON NATURAL FAMILY PLANNING, p. 171-180.

Durbin, J.; 1954, "Errors in Variables", REVIEW OF THE INSTITUTE OF INTERNATIONAL STATISTICS 22:23-32.

Ertel, J.E. and E.B. Fowlkes; 1976, "Some Algorithms for Linear Spline and Piecewise Multiple Linear Regression", JASA 71:640-648.

Farley, J.U. and M.J. Hinich; 1970, "A Test for a Shifting Slope Coefficient in a Linear Model", JASA 65(331):1320-1329.

Feder, P.I.; 1975, "On the Asymptotic Distribution Theory in Segmented Regression Problems -- Identified Case", THE ANNALS OF STATISTICS 3(1):49-83.

Ferin, J.; 1947, BRUX. MED. 27:2786.

Fuller, W.A.; 1969, "Grafted Polynomials as Approximating Functions", THE AUSTRALIAN JOURNAL OF AGRICULTURAL ECONOMICS 13:35-46.

Gallant, A.R. and W.A. Fuller; 1973, "Fitting Segmented Polynomial

Regression Models Whose Join Points Have to be Estimated", JASA 68: 144-147.

Hansen, M.H., W.N. Hurwitz, and W.G. Madow; 1953, SAMPLE SURVEY METHODS AND THEORY, Vol. 1, Wiley and Sons, Inc., New York.

Hartley, H.O., J.N.K. Rao, and Gerace Kiefer; 1969, "Variance Estimation with One Unit per Stratum", JASA 64:841-851.

Hasselblad, V., J.P. Creason, and W.C. Nelson; 1976, "Regression Using Hockey Stick Functions", ENVIRONMENTAL HEALTH EFFECTS RESEARCH SERIES EPA-600/1-76-024, National Technical Information Service, Springfield, VA.

Hawkins, D.M.; 1976, "Point Estimation of the Parameters of Piecewise Regression Models", APPLIED STATISTICS 25(1):51-57.

Hillebrand, W.; 1971, JOURNAL OF BIOSOCIAL SCIENCE 3:331.

Hinkley, D.V.; 1969, "Inference About the Intersection in Two-Phase Regression", BIOMETRIKA 56(3):495-504.

———; 1970, "Inference About the Change-Point in a Sequence of Random Variables", BIOMETRIKA 57(1):1-17.

Housner, G.W., and J.F. Brennan; 1949, "The Estimation of Linear Trends", ANNALS OF MATHEMATICAL STATISTICS 19:380-388.

Hudson, D.J.; 1966, "Fitting Segmented Curves Whose Join Points Have to be Estimated", JASA 61:1097-1129.

Jacobi, M.; 1876, "The Question of Rest for Women During Menstruation", BOYLSTON PRIZE, Putnam's Sons, New York.

Johnston, James; 1972, ECONOMETRIC METHODS, 2nd edition, McGraw Hill.

Lerman, P.M.; 1980, "Fitting Segmented Regression Models by Grid Search", APPLIED STATISTICS 29(1):77-84.

Lindley, D.V.: 1947, "Regression Lines and the Linear Functional Relationship", SUPPLEMENT TO JRSS 9:218-244.

Madansky, Albert; 1959, "The Fitting of Straight Lines When Both Variables Are Subject to Error", JASA 54:173-205.

Magoun, A.D.; 1978, "A Modification of Hudson's Technique for PWLR Models", Unpublished Dissertation, U. of Southwestern Louisiana.

Marshall, J.; 1968, "A Field Trial of the Basal Body Temperature Method of Regulating Birth", LANCET 2:8-10.

Martin, P.L.; 1943, "Detection of Ovulation by the Basal Temperature Curve With Correlating Endometrial Studies", AMERICAN JOURNAL OF

OBSTETRICS AND GYNECOLOGY 46:53-62.

McGee, V.E. and W.T. Carleton; 1970, "Piecewise Regression", JASA 65: 1109-1124.

Mellits, E.D.; 1968, "Statistical Methods", p. 19-38 in HUMAN GROWTH: CELL GROWTH, ENERGY, INTELLIGENCE, AND BODY COMPOSITION, ed. D.B. Cheek, Lea and Febiger, Philadelphia.

Monti, K.L., G.G. Koch, and J. Sawyer; 1978, "An Application of Segmented Linear Regression Models to the Analysis of Data From a Cross-Sectional Growth Experiment", Institute of Statistics, UNC, Mimeo Series No. 1200.

Moran, P.A.P.; 1971, "Estimating Structural and Functional Relationships", JOURNAL OF MULTIVARIATE ANALYSIS 1:232-255.

Nair, K.R. and K.S. Banerjee; 1943, "Note on Fitting of Straight Lines if Both Variables Are Subject to Error", SANKHYA 6:331.

Neter, J. and W. Wasserman; 1974, APPLIED LINEAR STATISTICAL MODELS, Richard D. Irwin, Inc., Homewood, IL.

Robison, D.E.; 1964, "Estimates for the Points of Intersection of Two Polynomial Regressions", JASA 59:214-224.

Royston, J.P. and R.M. Abrams; 1980, "An Objective Method for Detecting the Shift in BBT in Women", BIOMETRICS 36:217-224.

Squire, W.; 1868, TRANS. OBSTET. SOC. LONDON (9), p. 129.

Quandt, R.E.; 1958, "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes", JASA 55:873-880.

———; 1960, "Tests of the Hypothesis That a Linear Regression Obeys Two Separate Regimes", JASA 55:324-330.

Singpurwalla, N.D.; 1974, "Estimation of the Join Point in a Heteroscedastic Regression Model Arising in Accelerated Life Tests", COMMUNICATIONS IN STATISTICS 3(9):853-863.

Smith, Patricia L.; 1979, "Splines as a Useful and Convenient Statistical Tool", THE AMERICAN STATISTICIAN 33(2):57-62.

Thiel, H. and J. Van Yzeven; 1956, "On the Efficiency of Wald's Method of Fitting Straight Lines", REVUE INST. INT. DE STAT. 24:17-26.

Tukey, J.W.; 1951, "Components in Regression", BIOMETRICS 7:33-69.

Van de Velde, T.H.; 1928, Die Vollkommene Ehe, Eine Studie, UBER IHR PHYSIOLOGIE UND TECHNIK, Leipzig.

Vollman, R.F.; 1977, THE MENSTRUAL CYCLE, W.B. Saunders Co., Philadelphia.

Wade, M.E., et al.; 1981, "A Randomized Prospective Study of the Use-Effectiveness of Two Methods of Natural Family Planning", AMERICAN JOURNAL OF OBSTETRICS AND GYNECOLOGY 141(4):368-376.

Wald, A.; 1940, "The Fitting of Straight Lines if Both Variables Are Subject to Error", ANNALS OF MATHEMATICAL STATISTICS 11:284-300.