AN ABSTRACT OF THE THESIS OF

JAMES CORNELIUS DALY (Name) for the DOCTOR OF PHILOSOPHY (Degree)

in STATISTICS (Major) presented on August 31, 1973 (Date)

Title: A BAYESIAN APPROACH TO TWO-PHASE REGRESSION

Abstract approved: Redacted for privacy

Donald Guthrie

A Bayesian approach to the analysis of a two-phase linear regression model is given. It is assumed that the regression model is continuous at the change point. The likelihood function is expressed in a form which explicitly contains the continuity restriction. The natural conjugate prior distribution for the likelihood function is used, and the form of the prior constrained mean vector and dispersion matrix is developed for the situation where prior knowledge only exists on an unconstrained model.

The situation where the join point is unknown is approached by discretizing the possible values of the join point, and assuming a discrete prior distribution for the join point. The marginal posterior probability of the join point is determined, and the various posterior conditional and marginal distributions of the unknown parameters are shown. Numerical examples are considered which show the similarities and differences which exist between the Bayesian approach and the more common least squares approach. The situation of vague prior knowledge is briefly considered.

A Bayesian Approach to Two-Phase Regression

by

James Cornelius Daly

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

June 1974

APPROVED:

Redacted for privacy

Professor of Statistics
Adjunct Professor of Psychiatry
University of California, Los Angeles

in charge of major

Redacted for privacy

Chairman of Department of Statistics

Redacted for privacy

Dean of Graduate School

Date thesis is presented August 31, 1973

Typed by Clover Redfern for James Cornelius Daly

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

A BAYESIAN APPROACH TO TWO-PHASE REGRESSION

1. INTRODUCTION

1.1 Statement of the Problem

The subject of this dissertation is a Bayesian approach to two-phase regression. In two-phase regression, the dependent variable is seen as having one regression relationship with the independent variables over a certain segment of the range of the independent variables, and a different regression relationship with the independent variables over a second segment of the range of the independent variables. In the paper we are principally concerned with the case where there is one independent variable, both regression relationships are linear and the slope and intercept parameters of the two regressions are distinct.

Two-phase regression has applications in many fields. Examples of situations where the model is appropriate appear in the physical sciences (1), in agriculture (17), and in econometrics (13, 11). One simple-minded example of a two-phase regression problem arising in agriculture is the relationship of weight gain of cows to the amount of feed for the cows. If it is assumed that the weight gain per cow during a certain period is a linear function of the amount of feed available per cow, a two-phase regression model can be used. The

regression function has a positive slope until the amount of feed available reaches a point at which the weight gain per cow slows down. In other words, after the independent variable, namely the amount of feed available per cow, has reached a certain point, say $x(0)$, the slope of the regression function will level off. Two-phase regression models consist of two line segments defined over two distinct and adjacent regimes of the independent variable. The first regime of this example consists of all amounts of feed per cow greater than 0 and less than the value of $x(0)$. The second regime consists of all quantities of feed per cow that are greater than $x(0)$. The value $x(0)$ is called the change point of the two-phase regression model, indicating the value of the independent variable at which the slope parameter of the regression model changes values. In this example the main points of interest are estimation of the slope and intercept parameters of the line segment defined over the first regime, and the estimation of the change point. Because of the form of this example, any inference on the parameters of the regression line for the second regime is probably unnecessary. In most situations however, inference on this parameter is a significant part of the statistical analysis.

The situation studied in this paper may be more formally stated as follows: Let Y, the dependent variable, have a regression relationship with X, the independent variable, where we assume that X is measured without error. We assume the regression model is

$$(1.1) \qquad Y = \alpha_0 + \alpha_1 X + \epsilon, \quad \text{when} \quad X \leq \gamma,$$

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \text{when} \quad X > \gamma.$$

ϵ is assumed to be a normally distributed random variable with mean 0 and variance σ^2.

For this model α_0, α_1, β_0, β_1, γ and σ^2 are all assumed to be unknown parameters, unless otherwise stated. The parameter γ denotes the change point $x(0)$. γ will be used throughout the remainder of this paper.

In most cases the model should be continuous at the change point. This restriction insures that the regression line does not have an abrupt jump as the regimes change. Thus, the restriction

$$(1.2) \qquad \alpha_0 + \alpha_1 \gamma = \beta_0 + \beta_1 \gamma$$

is imposed on the model, making the change point the abscissa of the join point of the two regression lines. Except where confusion might arise, γ will be referred to as the join point of the two-phase regression model with the above continuity condition.

In the unrestricted model given by Equation (1.1), there are six unknown parameters. When the continuity condition is imposed, five of these parameters must be estimated; the estimate of the sixth parameter determined by the other estimates. The major unique problem involved in working with a two-phase regression model is not

inference about the two slopes and intercepts, but the inference concerning the location of the change point or join point.

1.2 Previous Approaches

We will now look at some of the approaches to the two-phase regression model that have been considered in the statistical literature. The approaches considered are not meant to represent all the approaches contained in the literature, but do give a good indication of how the analysis has developed. A look at these approaches will show the various problems encountered when using the different statistical techniques, and give some rationale for the Bayesian approach that will be used in this paper.

Assume the model stated in Equation (1.1). Denote a sample of size n from an experiment for which the two-phase regression model is applicable as n pairs (x_i, y_i), $i = 1, \ldots, n$, where the observations are ordered on the x's such that for i less than j implies that x_i is less than or equal to x_j, for all values of i and j from 1 to n. Also, assume that there are at least six pairs of observations in the sample, and that at least five of the x-values are distinct. If the continuity condition is imposed, the necessary sizes reduce to five and four, respectively. The above assumptions are necessary in order to have at least one situation in which all parameters can be estimated. Let the interval $I(i)$,

$i = 2, 3, \ldots, n-2$ be defined by

$$(1.3) \qquad I(i) = \{x \mid x_i \leq x < x_{i+1}\} .$$

If one assumes that the change point (or join point) is such that $x_j \leq \gamma < x_{j+1}$, then $I(j)$ will be called the interval of interest. If the data comes from a population to which the two-phase regression model applies, then

$$\gamma \in \bigcup_{i=1}^{n-1} I(i) = I .$$

The general restriction on the number of points in each regime is that if $\gamma \in I$, then there must be at least two distinct sample values of x less than or equal to γ, and at least two distinct sample values of x greater than γ. For the sake of convenience of notation and without any loss of generality, we assume that if

$$\gamma \in \bigcup_{i=1}^{n-1} I(i), \quad \text{then} \quad \gamma \in \bigcup_{i=2}^{n-2} I(i) ,$$

and $x_1 < x_2$ and $x_{n-1} < x_n$. Given this formulation and set of assumptions for the two-phase regression model, we will now look at the development of the statistical methods for the problem.

1.2.1 Maximum Likelihood Estimation

Quandt (13, 14) was one of the first people to consider the two-phase regression estimation problem. The model used in his analysis is somewhat different from that considered in (1.1), and is also different from most other models which will be discussed. The difference lies in how the two regimes are defined. Quandt assumed that some variable not included in the regression model determines the two regimes. An obvious example is the time or order in which the observations are taken, where time or order is not the independent variable in the regression equation. Because of this, his two regimes are not necessarily continuous functions of X, as is the case in our model given by (1.1), and all other models we will consider. Quandt also allowed the error ϵ to have a different variance in the two regimes, thus introducing another parameter. Because of his definition of the two regimes, the continuity condition given by (1.2) makes no sense in most cases.

Quandt's estimation procedure gives estimates of the two slope parameters, the two intercept parameters, the two variances, and the interval between which two observations the change most likely takes place. The observations are assumed ordered with respect to the outside variable defining the two regimes. The method of estimation is the maximum likelihood method and is based on finding $n-3$ different

sets of "conditional" maximum likelihood estimates and then choosing the best of these sets. The method starts by assuming a certain interval, say $I(j)$, is the interval in which the true change point lies. For this situation get the maximum likelihood estimates for the parameters of the two regression line segments. These estimates are of the same form as the estimates obtained in simple linear regression, given that the data in the two regimes is analyzed separately. With these estimates the likelihood function is evaluated for $I(j)$. The above operation is carried out for each interval as j ranges from 2 to $n-2$. The unconditional estimate of the interval in which the change occurs is that interval whose "conditional" estimates yield the largest value of the likelihood function. Call this estimate $\hat{I}$. The maximum likelihood estimates for the slopes, intercepts, and variances are simply the estimates derived when it is assumed that the interval $\hat{I}$ contains the true value of the change point. This procedure must be carried out for all $n-3$ intervals, since Quandt demonstrated that the likelihood function taken as a function of the interval of interest need not be a unimodal function.

Although Quandt's model is different from that given by (1.1), many people who assume a model such as (1.1) use a method similar to Quandt's to estimate their unknown parameters. If this is done, the estimates derived can lead to some strange results. For one thing, the two estimated regression lines need not intersect within the

interval of interest, thus giving us a positive or negative jump of the regression function somewhere within this interval. In many cases this result is very difficult to interpret. However, Quandt's method applied to (1.1) is fairly straightforward, and seems to enjoy a certain amount of popularity.

Sprent (17) and Robison (16) also consider maximum likelihood methods of estimation for the two-phase regression model. Sprent was principally concerned with developing tests for many different hypotheses about the parameters of the model. Among the hypotheses which he considered are the following: two lines of different slope meeting at some known point, two parallel lines having different values for the intercept parameters defined over the two regimes, and the case where the second line has infinite slope. The error rates of the tests derived are highly dependent on knowing the interval in which the change point occurs. The tests will not be discussed further in this paper.

Robison considered a model more general than that given by (1.1). He considered polynomial functions of one independent variable defined over the two regimes. These two polynomial functions need not necessarily be of the same degree. In his estimation procedure, Robison derives "conditional" maximum likelihood estimates for each potential "interval of interest," similar to Quandt, but after finding the estimates of the parameters for the two lines, he imposes the

constraint that the two lines must intersect in the "interval of interest." After finding the estimates for each interval, the method chooses the set of estimates that maximize the likelihood function, provided the estimate of the joint point falls in the interval being considered. When using this technique it is possible to have cases where no admissible estimates of the join point exist, and also there are cases where multiple estimates of the join point, for a given interval, exist. In this context, an admissible estimate of the join point is an estimate that satisfies the condition that the join point lies within a given interval. Robison gives heuristic techniques to alleviate these problems regarding admissible estimates. Fuller (6), and Gallant and Fuller (7) have done more recent work where the functions can be polynomial functions.

1.2.2 Hudson's Least Squares Approach

The most systematic approach to the analysis of the general two-phase regression model is due to Hudson (10). His approach lets the two functions be functions of any known form of one independent variable. This discussion will be limited to his treatment of the two-phase simple linear regression model. In this case the model assumed is the same as that given by (1.1) with continuity condition given by (1.2). The method of estimation used is the method of least squares. Hudson's technique looks at the possibility of the join point

lying in each of the possible intervals, and then chooses the interval, or equivalently the join point, which minimizes the residual sum of squares. If the errors are normally distributed, the estimates obtained by this method are the maximum likelihood estimates. The importance of Hudson's method is not the least squares approach which he uses, but rather the way in which he finds admissible least squares estimates.

Let $\hat{\alpha}_{0i}, \hat{\alpha}_{1i}$ be least squares estimates of the intercept and slope parameters of a linear regression line based on the first i points of our data. Let $\hat{\beta}_{0i}, \hat{\beta}_{1i}$ be the corresponding least squares estimates for the second straight line segment based on the last $n-i$ data points. These four estimates are called the naive least squares estimates based upon the condition that the true join point lies in the interval $I(i)$. Repeat the above for all i as i ranges from 2 to $n-2$. Let

$$\hat{\gamma}_i = \frac{(\hat{\alpha}_{0i} - \hat{\beta}_{0i})}{(\hat{\beta}_{1i} - \hat{\alpha}_{1i})}, \qquad i = 2, \ldots, n-2. \tag{1.4}$$

For every i where $\hat{\gamma}_i \in I(i)$, $\hat{\gamma}_i$ is an admissible estimate of the join point, and one calculates the Res. S.S. associated with the set of parameter estimates $(\hat{\alpha}_{0i}, \hat{\alpha}_{1i}, \hat{\beta}_{0i}, \hat{\beta}_{1i}, \hat{\gamma}_i)$. For every interval where $\hat{\gamma}_i \notin I(i)$, an admissible estimate for this interval is obtained. Hudson cites a theorem by McLaren (12) which says that when

$\hat{\gamma}_i \notin I(i)$, the best restricted estimates in a least squares sense are achieved when the join point is set equal to one of the two end points of $I(i)$, x_i or x_{i+1}, usually the endpoint which is closest to the unrestricted estimate of γ. Although we have previously defined $I(i)$ as a half open interval, the inclusion of the upper end point in this technique is necessary. Hudson develops techniques so that any end point is not considered more than once. McLaren's theorem is mentioned in other papers, but the conditions in the different papers necessary for the theorem to apply are not equivalent. It appears that if the unrestricted estimate of the join point for a given interval is outside that interval but not too far away from one of the endpoints, then the results of the theorem will hold. The theorem is used in the following way. Suppose that for some j, $\hat{\gamma}_j \notin I(j)$. Proceed by finding constrained least squares estimates for the two slope and two intercept parameters for the case where we have the restriction

$$x_j = (\alpha_0 - \beta_0)/(\beta_1 - \alpha_1) ,$$

and then find the estimates for the case where

$$x_{j+1} = (\alpha_0 - \beta_0)/(\beta_1 - \alpha_1) .$$

Whichever of these two sets of restricted estimates yields the smallest Res. S.S. is the set which will be called the least squares

estimates for interval I(j), and the endpoint which yields this set of estimates is our estimate of the join point for I(j). This procedure is carried out for all intervals that do not have an admissible unconstrained estimate of the join point. The "optimal" estimates for the model are simply the set of admissible estimates, constrained or unconstrained, that yield the minimum Res. S.S. Hudson describes an algorithm which yields the best estimates in the least number of steps. The method outlined above and a more completely developed method in Hudson's paper is the method by which a two-phase regression model which is continuous at the join point is usually estimated.

The method of Hudson is very effective for point estimation of the parameters in a two-phase regression model. However, it is strictly concerned with point estimation. Hinkley (8, 9) worked on interval estimation problems and hypothesis testing for the two-phase regression model. His inference was based on maximum likelihood estimates for models with normal errors. Relying on the fact that, except in rare cases, the maximum likelihood estimates of the two slopes and the join point are asymptotically normally distributed, Hinkley developed large sample tests and confidence intervals for the various parameters. Since the results are based on large sample theory, they are only approximate. In Monte Carlo studies conducted by Hinkley, the distributions of the estimates of the slopes converged to their theoretical limiting distributions quite quickly,

while the distribution of the estimates of the join point converged slowly to normality. Because of this, Hinkley derived an alternate asymptotic distribution for the estimate of the join point which gives a better fit to the sampling distribution for moderate sample sizes. However, for small samples the distribution deviated severely from the empirical distribution derived in the Monte Carlo study. One especially interesting result of his small sample study is that for small samples the maximum likelihood technique imposes a bias in $(\beta_1 - \alpha_1)$, thus imposing a bias into the estimate of the join point.

1.2.3 Other Estimation Techniques

Another significant contribution is due to Bacon and Watts (1). Their approach to the problem was based on a Bayesian method of analysis. The model which they considered is not the same as that given by Equation (1.1). Because their principal objective was to make inference about the transition between the two straight line segments, they put a parameter of curvature into the model, and were primarily concerned with estimating this new parameter and the value of the join point. The introduction of the parameter of curvature means that the transition from one straight line segment to a second straight line segment need not be abrupt, but can be the result of a gentle curve occurring within the interval of change. In some ways this method is actually looking at a three-phase regression model,

where the first and third phases are straight line segments, while the second, connecting phase is of the form of a smooth, continuous curve. Bacon and Watts assume noninformative priors on all unknown parameters, and they based their estimates on the model of the joint marginal posterior distribution of the parameter of curvature and the join point. The approach to be developed in this paper is also of a Bayesian nature, but the parameter of curvature is not included. Much more attention is given to the form of the prior distribution. Inference on parameters other than the join point is also included in the analysis.

Some recent approaches to the two-phase regression problem, or more generally, to the multiple-phase regression problem use methods of solution not commonly associated with regression problems. Bellman and Roth (2) use a dynamic programming approach to arrive at an allocation of points to the different regimes that is optimal in some sense. McGee and Carleton (11) use a cluster analysis technique to arrive at the optimal allocation of the data points to the different regimes. In these methods, the regression functions are assumed linear over the various regimes. Bellman, with Kashef and Vasudevan (3), has also used dynamic programming to consider the case where the functions defined within the various intervals are all cubic polynomials whose values are known at the various endpoints. All of these papers are primarily pragmatic approaches to the

problem, and we will not discuss them further in this paper.

1.2.4 Bayesian Approach of This Paper

All of the preceding approaches have their value and, for the most part, answered the questions which they intended to answer. Why then is there any need to consider another approach to the two-phase simple linear regression model when most current papers are concerned with more complex models? There are two primary reasons for further consideration of this problem, especially with a Bayesian approach. In the first place despite all of the work which has been done on the problem, classical statistical methods still have inferential problems with a two-phase regression model when the interval in which the join point occurs is unknown. Theoretically, the Bayesian approach is able to handle this problem much easier. This can be done by merely introducing another unknown parameter into the framework of the problem.

However, by far the biggest advantage of the Bayesian approach is that it takes advantage of prior knowledge of the form of the model under consideration. As is the case in many Bayesian problems, an experimenter might be unable to specify prior distributions for all of the unknown parameters, but it is extremely likely that if he suspects his model is a two-phase linear model, he does have some belief about the location of the join point, or the possible values of the slope

parameters. The Bayesian approach developed in this paper can take full advantage of this knowledge. The method of Bacon and Watts, although Bayesian in outlook and original in its approach, uses a noninformative prior distribution, and hence loses some of the versatility inherent in the Bayesian method of analysis.

In Chapter 2 the Bayesian approach for the two-phase regression problem with known join point is developed, in Chapter 3 the Bayesian approach for the situation where the join point is unknown is studied, and in Chapter 4 the Bayesian approach for our model with a vague prior is studied.

1.3 Example

Some people might question the use of regression models which consist of more than one submodel, insisting that one, slightly more complicated model gives a better representation of the actual model which generated the data. At this time we will present an example consisting of data which, although artificially generated, shows the advantage of using two-phase regression models. With this example we will illustrate the usual approach, due to Hudson, by which the parameters of a two-phase linear regression model are usually estimated.

The data used in this example was generated from the following model:

(1.5) $$Y = 2.5 + 2X + \epsilon, \quad \text{when} \quad X \leq 6.5,$$

$$Y = 12.25 + .5X + \epsilon, \quad \text{when} \quad X > 6.5.$$

The errors are assumed to be normally distributed with mean zero and variance 1. The ten generated observations are

X	1	2	3	4	5	6	7	8	9	10
Y	3.18	6.80	6.73	10.02	13.41	15.57	16.55	16.31	15.03	16.45

The data is plotted in Figure 1.

We will analyze this data as if it has come from one of three possible models: (I), a simple linear regression function defined over all possible values of X; (II), a polynomial function defined over all possible values of X; and, (III), a two-phase linear regression model with unknown join point.

(I) Simple linear regression model. Our model equation in this case is

$$Y_i = \alpha_0 + \alpha_1 X_i + \epsilon_i, \quad i = 1, \ldots, 10,$$

where $E(\epsilon_i) = 0$, and $Var(\epsilon_i) = \sigma^2$. The least squares estimates of the parameters for this model for the sample data is given in Figure 2.

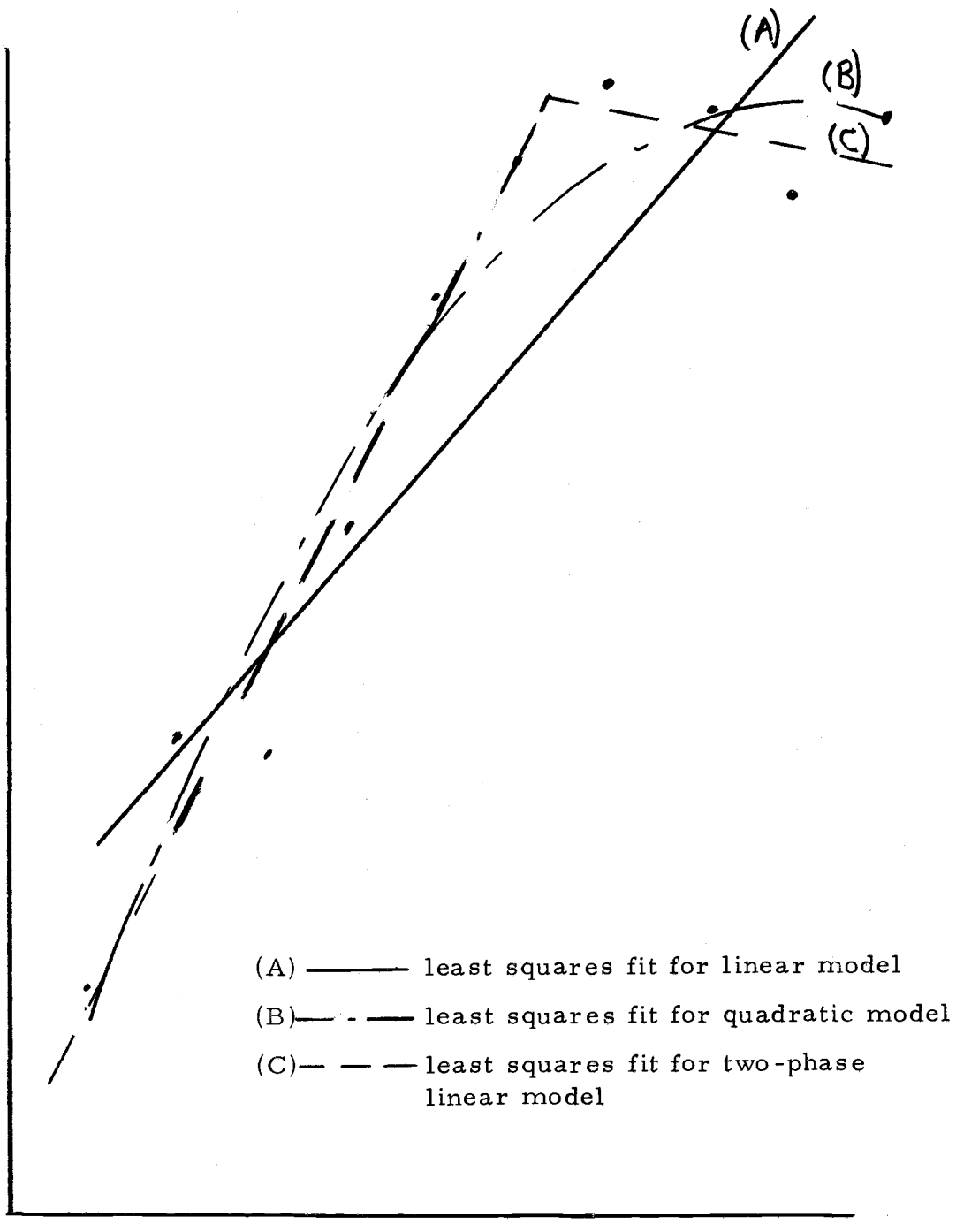

Figure 1. Plot of data points for example 1.4.

Parameter	α_0	α_1	σ^2
Estimate	3.78	1.49	4.38

Figure 2. Least squares estimates for a simple linear model.

(II) Polynomial regression model. The model equation is

$$Y_i = \alpha_0 + \alpha_1 X_i + \ldots + \alpha_k X_i^k + \epsilon_i, \quad i = 1, \ldots, 10,$$

where $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. We will consider four polynomial functions, as k ranges from 2 to 5. Table 1.1 contains the least squares estimates of the parameters for the different degree polynomials. By studying the last row of the table, we see that the estimate for the fifth degree polynomial is much smaller than the estimates of variance for the smaller degree polynomials. However, using a fifth degree polynomial to explain ten observations is a highly debatable practice.

(III) Two-phase linear regression model. In this situation the model equation is

$$Y_i = \alpha_0 + \alpha_1 X_i + \epsilon_i, \quad \text{when} \quad X_i \leq \gamma,$$
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \text{when} \quad X_i > \gamma,$$

where the errors have the same requirements as (I) and (II). In the

analysis we will only consider those values of X that lie in the closed interval $[5, 8]$ as potential values for the join point, since by a visual inspection of Figure 1 values in this range are much more likely than any of the other possible values. Therefore, we consider $I(5)$, $I(6)$, and $I(7)$, plus, if necessary, the value $x = 8$. By studying the entries in Figure 3, we see that the unconstrained estimates of the join point for $I(5)$ and $I(7)$ are both inadmissible; that is, $4.3 \notin I(5)$, and $6.36 \notin I(7)$. Therefore, to find admissible estimates for these intervals, we must consider the estimates which are constrained to meet at the end points of each of these two intervals. However, if we look at the last column of Figure 3, we see that the Res. S.S. for $I(6)$ is less than the Res. S.S. for each of the other two intervals. For any given interval, the Res. S.S. for the unconstrained estimates is less than or equal to the Res. S.S. for the estimates constrained to meet at either of the end points of that interval. Therefore, the interval which gives us the minimum Res. S.S. in our example is $I(6)$. The estimates are listed in Figure 4.

How do the various models compare? Model I, with $\sigma^2 = 4.38$ is easily dominated by both Model II and Model III, if we use minimum residual mean squares as our criterion in choosing a model. These results are to be expected, since both of the last two models have additional parameters to take care of the definite non-linearity of the

Table 1.1. Least squares estimates of parameters for various degree polynomials.

Parameter	k = Degree of Polynomial			
	2	3	4	5
a_0	-.987	1.038	5.17	-5.60
a_1	3.88	2.083	-3.22	14.54
a_2	-.22	.172	2.11	-7.21
a_3		-.024	- .29	1.79
a_4			.01	-.19
a_5				.01
σ^2	1.47	1.43	1.24	.44

Interval	Estimate of join point	Res. S.S.
I(5) = [5, 6)	4.316	5.275
I(6) = [6, 7)	6.433	5.035
I(7) = [7, 8)	6.361	5.688

Figure 3. Estimates of join points and Res. S.S. for two-phase linear regression.

Parameter	γ	a_0	a_1	β_0	β_1	σ^2
Estimate	6.433	.778	2.43	17.43	-.158	.848

Figure 4. Optimal least squares estimates for a two-phase regression problem.

data. By comparing Table 1.1 and Figure 4, we see that the two-phase regression model fits better than polynomial functions up to and including the fourth degree. The fifth degree polynomial does have the smallest residual variance among the models considered, but the two-phase linear regression model is obviously much easier to interpret.

This example illustrates some of the possible power of the two-phase linear regression model. At the end of Chapter 3, we will again study the data of this example, showing how a problem such as this can be analyzed by the methods developed in this paper.

2. A BAYESIAN PROCEDURE FOR THE ANALYSIS OF A TWO-PHASE LINEAR REGRESSION MODEL: I. JOIN POINT KNOWN

In any problem in which Bayesian inference is used, the statistician has three parts to consider: the prior distribution of the unknown parameters; the likelihood function arising from the sample; and the posterior distribution of the unknown parameters. Since the posterior distribution depends on the prior distribution and the likelihood function, care should be taken in the development of the appropriate expression of the likelihood function and the choice of the prior distribution.

One of the most tractable and rich classes of prior distributions is the conjugate prior distributions. This class of prior distributions insures that the posterior distribution belongs to the same family as the prior distribution. This and other properties of conjugate prior distributions can be found in Raiffa and Schlaffer (15). In many instances in order to identify the conjugate prior distribution for a given problem, the likelihood function of the observations must be expressed in a form different than the form used in classical statistical analysis. The analysis developed in this paper will use a family of prior distributions that is conjugate to the likelihood function, and because of this fact, some attention must be given the development of the appropriate expression of the likelihood function. In this chapter the analysis of the two-phase linear regression model where the join

point is known will be developed. The Bayesian analysis of the more interesting and practical problem of the unknown join point will be considered in the following chapter.

2.1 The Likelihood Function

Assume the continuous model given by (1.1) and (1.2), and assume a sample of size n, (x_i, y_i), $i = 1, \ldots, n$, where $x_i \leq x_m$ for $i < j$. Further assume that the known value of the join point γ is contained in the half open interval, $I(r)$, where r is some integer greater than or equal to 2, less than or equal to $n-2$, and

$$I(r) = \{x \mid x_r \leq x < x_{t+1}\} .$$

The two-phase linear regression model for this data can be written as

(2.1) $$\underline{Y} = \underline{X\phi} + \underline{\epsilon}$$

where

(2.2) $$\underline{Y}: n \times 1 = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X: n \times 4 = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ \vdots & & & \vdots \\ 1 & x_r & 0 & 0 \\ 0 & 0 & 1 & x_{r+1} \\ \vdots & & & \vdots \\ 0 & 0 & 1 & x_n \end{bmatrix}, \quad \underline{\phi}: 4 \times 1 = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \beta_0 \\ \beta_1 \end{bmatrix},$$

and $\underline{\epsilon}$ is an $(n \times 1)$ column vector for which it is assumed

$$\underline{\epsilon} \sim N_n(\underline{0}, \sigma^2 I_n) .$$

Given this structure, the likelihood of the observation, denoted $\ell(\underline{\phi}, \sigma ; X, Y)$, is proportional to

$$\sigma^{-n} \exp\{-(2\sigma^2)^{-1}(\underline{Y}-X\underline{\phi})'(\underline{Y}-X\underline{\phi})\} , \tag{2.3}$$

with the parameter vector $\underline{\phi}$ restricted by the condition $\underline{\delta}'(\gamma)\underline{\phi} = 0$, where

$$\underline{\delta}'(\gamma) = (1 \quad \gamma \quad -1 \quad -\gamma) . \tag{2.4}$$

This restriction insures that the regression lines intersect at a point where the abscissa is equal to γ. Identifying a conjugate prior for this problem is more straightforward if the above condition is contained explicitly in the likelihood function. To show this let $A(\gamma)$ be a 4×3 matrix such that

$$\underline{\delta}'(\gamma)A(\gamma) = \underline{0}' , \tag{2.5}$$

where $\underline{0}$ is a column null vector of size 3 and the column rank of the matrix A is equal to 3. A depends on the join point γ, but we will suppress the dependence on γ, and denote the matrix as A except where confusion might arise. Using this matrix, the likelihood

function denoted by (2.3) with the continuity restriction given by (1.2) can be reparameterized and written as

$$\ell(\underline{\theta}, \sigma; X, Y, A) \propto \sigma^{-n} \exp\{-(2\sigma^2)^{-1}(\underline{Y} - \underline{XA\theta})(\underline{Y} - \underline{XA\theta})\}$$

where $\underline{\theta}$ is such that

$$A\underline{\theta} = \underline{\phi} \,. \tag{2.6}$$

By using the above formulation the model now has three unrestricted parameters (contained in $\underline{\theta}$). The vector $A\underline{\theta}$ satisfies the condition that the two regression lines intersect at the join point, since

$$\underline{\delta}'(\gamma)\underline{\phi} = \underline{\delta}'(\gamma)A\underline{\theta} = 0 \,.$$

By letting $XA(\gamma) = W$ the notation can be further simplified, giving

$$\ell(\underline{\theta}, \sigma; W, Y) \propto \sigma^{-n} \exp\{-(2\sigma^2)^{-1}(\underline{Y} - \underline{W\theta})'(\underline{Y} - \underline{W\theta})\} \,. \tag{2.7}$$

It is obvious that the matrix W is also dependent on γ, but W will be used in place of $W(\gamma)$, except where confusion in notation might arise.

The form of the likelihood function can be modified further. In classical statistics, the sample is frequently reduced to sufficient statistics, when they exist. There exists a corresponding concept in Bayesian statistics. A function of the sample $\underline{Y}$, say $t(\underline{Y})$, is

sais to be Bayesian sufficient for the vector of parameters $\underline{\theta}$; if for any prior distribution on $\underline{\theta}$, say $P(\underline{\theta})$, the posterior distribution of $\underline{\theta}$ given $\underline{Y}$, $P'(\underline{\theta}/\underline{Y})$, is such that

$$P'(\underline{\theta}/\underline{Y}) = P'(\underline{\theta}/t(\underline{Y})).$$

Raiffa and Schlaiffer (15) show that classical sufficiency and Bayesian sufficiency are equivalent.

How does this change the likelihood function? $-2\sigma^2$ times the exponent of the likelihood function can be expressed as

$$(2.8) \qquad (Y-W\theta)'(Y-W\theta) = (\underline{Y}-W\underline{\hat{\theta}})'(\underline{Y}-W\underline{\hat{\theta}}) + (\underline{\theta}-\underline{\hat{\theta}})'W'W(\underline{\theta}-\underline{\hat{\theta}}) ,$$

where

$$(2.9) \qquad \underline{\hat{\theta}} = (W'W)^{-1}W'\underline{Y} .$$

From (2.8) by using the factorization theorem for sufficient statistics, we find that the set of statistics $((Y-W\underline{\hat{\theta}})'(Y-W\underline{\hat{\theta}}), \underline{\hat{\theta}})$ are jointly sufficient for the set of parameters $(\sigma^2, \underline{\theta})$ in the classical sense. Thus, the same set of statistics are Bayesian sufficient, and the likelihood function is, without any loss of information, proportional to

$$(2.10) \qquad \sigma^{-n}\exp\{-(2\sigma^2)^{-1}(\underline{Y}-W\underline{\hat{\theta}})(\underline{Y}-W\underline{\hat{\theta}})-(2\sigma^2)^{-1}(\underline{\theta}-\underline{\hat{\theta}})'W'W(\underline{\theta}-\underline{\hat{\theta}})\} .$$

This is the form of the likelihood function that will be principally used in the development of the Bayesian analysis.

In order to better understand any inference that is made on the parameters in the reparameterized model, the relationship between $\underline{\theta}$ and $\underline{\phi}$ must be considered. The vector of parameters $\underline{\theta}$ depends on the matrix A, and to give $\underline{\theta}$ a representation that is meaningful, the matrix A should be of a specific form. There are many possible transformations that satisfy the requirement given by (2.5). The matrix A that will be used in this Bayesian analysis is only one of many possibilities. Let

(2.11)
$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & -\gamma \\ 0 & 1 & 1 \end{bmatrix}.$$

This matrix satisfies the requirement $\underline{\delta}'(\gamma)A = \underline{0}'$ and has rank 3. If the designated matrix A has column rank less than 3, too many restrictions are being placed on the original vector of parameters $\underline{\phi}$, and the resulting parameter space over which the likelihood function is defined would be much too restrictive. Given the above matrix A, the following correspondence exists between $\underline{\phi}$ and $\underline{\theta}$.

$$\underline{\phi} = A\underline{\theta} \Rightarrow \begin{array}{ll} \alpha_0 = \phi_1 = \theta_1 & \beta_0 = \phi_3 = \theta_1 - \gamma\theta_3 \\ \alpha_1 = \phi_2 = \theta_2 & \beta_1 = \phi_4 = \theta_2 + \theta_3 . \end{array}$$

From this we see that

$$(2.12) \qquad \theta_1 = \alpha_0, \quad \theta_2 = \alpha_1, \quad \text{and} \quad \theta_3 = \beta_1 - \alpha_1 .$$

Thus, the three parameters of interest in the new formulation are the y-intercept and the slope of the first segment of the regression model, and the difference between the slopes of the second line segment and the first line segment. Given that the abscissa of the join point, namely γ, is known, inference about the three parameters in $\underline{\theta}$ and σ^2, is all that need be considered. The estimate of the intercept for the second line segment depends on the value of γ, and the estimates of θ_1 and θ_3. It should be kept in mind that the likelihood function that is used in this analysis is explicitly dependent on the join point through the matrix W, and any change in the value of γ will change the likelihood function.

2.2 Prior Distribution

In determining the prior distribution for a two-phase regression model, we first look at the case where the variance is assumed known, and find the conditional prior distribution for a particular value of the variance. Then we derive the joint prior distribution for $\underline{\theta}$ and h. As set forth in the first part of this chapter, the family of priors used in this analysis is the natural conjugate family for the likelihood function.

2.2.1 Prior Distribution, Variance Known

The conjugate prior for the parameter vector $\underline{\phi}$ in a multiple regression with uncorrelated normally distributed errors is (Raiffa and Schlaiffer)

$$N_4(\underline{\mu}_1(\underline{\phi}), \Sigma_1(\phi)),$$

the multivariate normal distribution with mean vector $\underline{\mu}_1(\underline{\phi})$, and dispersion matrix $h^{-1}\Sigma_1(\phi)$, where $h = \sigma^{-2}$.

If this prior distribution on the parameters is assumed when there is no constraint on the parameters, what form does the conjugate prior take when the continuity restriction is placed on the parameters, and we are working with transformations of original parameters? Let $\underline{\rho} = P\underline{\phi}$, where

$$P = \begin{bmatrix} I_4 \\ \hline \underline{\delta}'(\gamma) \end{bmatrix}, \tag{2.13}$$

I_4 is the 4-dimensional identify matrix, and $\underline{\delta}'(\gamma)$ is as defined in (2.4). It follows that the 5-dimensional vector $\underline{\rho}$ also has a multivariate normal distribution, namely,

$$\underline{\rho} \sim N_5(\underline{\mu}_1(\rho), h^{-1}\Sigma_1(\rho)) ,$$

where

$$\underline{\mu_1(\rho)} = P\underline{\mu_1(\phi)}, \quad \text{and} \quad \Sigma_1(\rho) = P\Sigma_1(\phi)P' .$$

In this formulation,

$$\underline{\rho}' = (\rho_1\rho_2\rho_3\rho_4\rho_5) = (\phi' : \underline{\delta'(\gamma)\phi}) .$$

Thus, the first four elements of $\underline{\rho}$ are the four elements of $\underline{\phi}$, and the fifth element of $\underline{\rho}$ is the linear combination of $\underline{\phi}$ which was constrained in the original formulation of the likelihood function.

Consider the conditional distribution of $\rho_1\rho_2\rho_3\rho_4/\rho_5$. By using the properties of conditional distributions of normal random variables it is readily seen that

$$(2.14) \qquad (\underline{\phi}/\underline{\delta'(\gamma)\phi}) \sim N\left[\begin{array}{l} (\underline{\mu}_1^{(1)}(\rho) + \Sigma_1(\rho)_{12}\Sigma_1^{-1}(\rho)_{22}(\rho_5 - \mu_1^{(2)}(\rho)), \\ h^{-1}(\Sigma_1(\rho)_{11} - \Sigma_1(\rho)_{12}\Sigma_1^{-1}(\rho)_{22}\Sigma_1(\rho)_{21}) \end{array}\right],$$

where

$$\mu_1^{(1)}(\rho) = \begin{bmatrix} \mu(\rho_1) \\ \mu(\rho_2) \\ \mu(\rho_3) \\ \mu(\rho_4) \end{bmatrix}, \quad \mu_1^{(2)}(\rho) = \mu(\rho_5) ,$$

and

(2.15)
$$\Sigma_1(\rho) = \left[\begin{array}{c|c} \underset{(4\times4)}{\Sigma_1(\rho)_{11}} & \underset{(4\times1)}{\Sigma_1(\rho)_{12}} \\ \hline \underset{(1\times4)}{\Sigma_1(\rho)_{21}} & \underset{(1\times1)}{\Sigma_1(\rho)_{22}} \end{array}\right]$$

Now, from the form of P we know that

$$\underline{\mu_1(\rho)} = P\underline{\mu_1(\phi)} = \left[\begin{array}{c} I_4 \\ \hline \underline{\delta'(\gamma)} \end{array}\right] \underline{\mu_1(\phi)} = \left[\begin{array}{c} \underline{\mu_1(\phi)} \\ \hline \underline{\delta'(\gamma)\ \mu_1(\phi)} \end{array}\right]$$

so

$$\underline{\mu_1^{(1)}(\rho)} = \underline{\mu_1(\phi)}, \quad \text{and} \quad \mu_1^{(2)}(\rho) = \underline{\delta'(\gamma)\ \mu_1(\phi)}\ .$$

Also,

(2.16)
$$\Sigma_1(\rho) = P(\Sigma_1(\phi))P' = \left[\begin{array}{c} I_4 \\ \hline \underline{\delta'(\gamma)} \end{array}\right] \Sigma_1(\phi)[I_4 \mid \delta(\gamma)]$$

$$= \left[\begin{array}{c|c} \underset{(4\times4)}{I_4\Sigma_1(\phi)I_4} & \underset{(4\times1)}{I_4\Sigma_1(\phi)\underline{\delta(\gamma)}} \\ \hline \underset{(1\times4)}{\underline{\delta'(\gamma)}\Sigma_1(\phi)I_4} & \underset{(1\times1)}{\underline{\delta'(\gamma)}\Sigma_1(\phi)\underline{\delta(\gamma)}} \end{array}\right]$$

$$= \left[\begin{array}{c|c} \Sigma_1(\phi) & \Sigma_1(\phi)\underline{\delta(\gamma)} \\ \hline \underline{\delta'(\gamma)}\Sigma_1(\phi) & \underline{\delta'(\gamma)}\Sigma_1(\phi)\underline{\delta(\gamma)} \end{array}\right]$$

Now, using the correspondence between (2.15) and (2.16), and the results on the mean vectors, (2.14) can be written as

$$(\underline{\phi}/\underline{\delta'(\gamma)\phi}) \sim N_4[\underline{\mu_1(\phi)}+\Sigma_1(\phi)\underline{\delta(\gamma)}\{\underline{\delta'(\gamma)}\Sigma_1(\phi)\underline{\delta(\gamma)}\}^{-1}(\underline{\delta'(\gamma)\phi}-\underline{\delta'(\gamma)\mu_1(\phi)}),$$

$$h^{-1}\{\Sigma_1(\phi)-\Sigma_1(\phi)\underline{\delta(\gamma)}\{\underline{\delta'(\gamma)}\Sigma_1(\phi)\underline{\delta(\gamma)}\}^{-1}\underline{\delta'(\gamma)}\Sigma_1(\phi)\}] .$$

Let

$$H(\gamma) = \Sigma_1(\phi)\underline{\delta(\gamma)}\{\underline{\delta'(\gamma)}\Sigma_1(\phi)\underline{\delta(\gamma)}\}^{-1}\underline{\delta'(\gamma)} . \tag{2.17}$$

If we consider the special case in which we are interested, namely $\underline{\delta'(\gamma)\phi} = 0$, we get

$$(\underline{\phi}/\underline{\delta'(\gamma)\phi} = 0) \sim N_4[(I-H(\gamma)\underline{\mu_1(\phi)}, h^{-1}((I-H(\gamma))\Sigma_1(\phi)] . \tag{2.18}$$

Equation (2.18) gives the conjugate distribution of the four original parameters of a two-phase regression model, namely the two slopes and intercepts under the restriction that the regression lines meet at a known join point. However, this is not the distribution with which we will work. This distribution is a singular 4-dimensional normal distribution since a restriction has been imposed on a linear combination of the four original "variables." Also, the likelihood function previously developed in (2.10) was formulated with a transformation of the original parameters. Therefore the distribution which we desire is the prior distribution for $\underline{\theta} = \underline{B\phi}$, where the distribution of $\underline{\phi}$ is conditioned on $\underline{\delta'(\gamma)\phi} = 0$, and

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} .$$

Finding the distribution of the above linear transformation of the vector $\underline{\phi}$ conditioned on $\underline{\delta'(\gamma)\phi} = 0$ results in the fact that

$$\underline{\theta} \sim N_3(\underline{\mu_1(\theta)}, h^{-1}\Sigma_1(\theta)) , \tag{2.19}$$

where

$$\underline{\mu_1(\theta)} = B(I-H(\gamma))\underline{\mu_1(\phi)} ,$$

$$\Sigma_1(\theta) = B(I-H(\gamma))\Sigma_1(\underline{\theta})B' ,$$

and

$$\underline{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_4 \end{pmatrix} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_4 - \phi_2 \end{pmatrix} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \beta_1 - \alpha_1 \end{pmatrix} .$$

$\underline{\theta}$ has the same form as the vector of parameters used in the last formulation of the likelihood function. Following the method by which the conjugate prior for $\underline{\theta}$ was derived, we are also assured that the distribution of $\underline{\theta}$ is such that the continuity condition is imposed on the mean vector and dispersion of $\underline{\theta}$. Assuming that the matrix $\Sigma_1(\phi)$ is nonsingular, the dispersion matrix $\Sigma_1(\theta)$ is also nonsingular, and we once again have a nonsingular normal distribution.

There is a slight possibility $\Sigma_1(\theta)$ will be singular. This situation will be discussed later.

When there is no possibility of misinterpretation, let

$$\underline{\mu_1} = \underline{\mu_1(\theta)}, \quad \text{and} \quad \Sigma_1 = \Sigma_1(\theta) .$$

Most of the following work will proceed with this notation. However, it should be realized that in terms of the original mean vector and dispersion matrix for our prior distribution, namely $\underline{\mu_1(\phi)}$ and $\Sigma_1(\phi)$, there exists the following correspondence;

$$(2.20) \qquad \underline{\mu_1} = B\left[(I-\Sigma_1(\phi)\underline{\delta(\gamma)}\{\underline{\delta'(\gamma)}\Sigma_1(\phi)\underline{\delta(\gamma)}\}^{-1}\underline{\delta'(\gamma)}\right]\underline{\mu_1(\phi)} ,$$

$$\Sigma_1 = B\left[(I-\Sigma_1(\phi)\underline{\delta(\gamma)}\{\underline{\delta'(\gamma)}\Sigma_1(\phi)\underline{\delta(\gamma)}\}^{-1}\underline{\delta'(\gamma)}\Sigma_1(\phi)\right]B' .$$

Using the simplified notation, (2.19) tells us that the prior density function of the vector of parameters $\underline{\theta}$, for a given join point γ, an original mean vector $\underline{\mu_1(\phi)}$, and dispersion matrix $\Sigma_1(\phi)$ is

$$(2.21) \quad f_1(\underline{\theta}/h, \gamma) = (2\pi)^{-3/2} h^{3/2} |\Sigma_1|^{-1/2} \exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu_1})'\Sigma_1^{-1}(\underline{\theta}-\underline{\mu_1})\} .$$

At this point some mention should be made of the method used to find the conjugate prior distribution for the vector of parameters $\underline{\theta}$. By its very nature, the kernel of the class of conjugate prior distributions for a certain likelihood function can be written down by

knowing the form of the likelihood function when it is expressed as a function of the sufficient statistics. Using the last form of our likelihood function (2.10) we see that the conjugate prior in a regression situation with known variance and three unknown parameters is a trivariate normal distribution. However, in the particular situation considered in this paper, if we had merely written down a trivariate normal distribution as our prior, we would have no idea what restrictions must be placed on the form of the mean vector and dispersion matrix of $\underline{\theta}$. By imposing the necessary restrictions on a 4-dimensional prior distribution, the restrictions on the prior parameters are explicit. This method gives the mean vector and dispersion matrix for the vector $\underline{\theta}$, for any arbitrary choice of $\underline{\mu_1(\phi)}$ and $\Sigma_1(\phi)$. Examples of how particular choices of $\underline{\mu_1(\phi)}$ and $\Sigma_1(\phi)$ are affected by the various transformations are given in Section 2.4, in connection with a numerical example.

2.2.2 Prior Distribution, Variance Unknown

In the previous section, the symbol h was introduced, where $h = \sigma^{-2}$. This parameter will be used in place of the variance. In Bayesian terminology, h is called the precision of the distribution. The interpretation of h is reciprocal to the interpretation of σ^2, i.e., a small value of h indicates a distribution with large spread, and a large value of h, similar to a small value of σ^2, indicates

a small spread.

Using this new notation, the likelihood function given in (2.10) is proportional to

$$h^{n/2} \exp\{-\frac{h}{2}(\underline{Y}-W\underline{\hat{\theta}})'(\underline{Y}-W\underline{\hat{\theta}}) - \frac{h}{2}(\underline{\theta}-\underline{\hat{\theta}})'W'W(\underline{\theta}-\underline{\hat{\theta}})\},$$

which can also be written as

$$h^{n/2} \exp\{-\frac{h}{2}(\underline{\theta}-\underline{\hat{\theta}})'W'W(\underline{\theta}-\underline{\hat{\theta}})\} \exp\{-\frac{h}{2}(n-3)s^2\}, \tag{2.22}$$

where

$$s^2 = (n-3)^{-1}(\underline{Y}-W\underline{\hat{\theta}})'(\underline{Y}-W\underline{\hat{\theta}})$$

is the unbiased estimator for σ^2 for a full rank linear model with three parameters. For the prior conjugate distribution when the precision, or variance, in unknown, a form of the gamma distribution will be used as the marginal conjugate prior for h. This density function $b_1(h)$ is

$$\frac{h^{\nu_1/2-1} \exp\{-h\nu_1 v_1/2\}(\nu_1 v_1/2)^{\nu_1/2}}{\Gamma(\nu_1/2)} \tag{2.23}$$

This distribution is related to the gamma by the following correspondence; $\alpha = (\nu_1/2)$, $\beta = (2/\nu_1 v_1)$. In using this prior distribution for h, it is assumed that h is independent of γ.

Since (2.23) is the marginal prior density function for h, combining it with the prior distribution of $\underline{\theta}$ for a fixed h, yields the joint prior density function of $\underline{\theta}$ and h. Thus, (2.21) derived in the previous section is combined with (2.23) to yield a joint conjugate prior density function for $\underline{\theta}$ and h, denoted $g_1(\underline{\theta}, h)$, proportional to

$$(2.24) \quad h^{p_1/2} |\Sigma_1|^{-1/2} \exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu}_1)'\Sigma_1^{-1}(\underline{\theta}-\underline{\mu}_1)\} \times h^{\nu_1/2} \exp\{-h\nu_1 v_1/2\}(\nu_1 v_1/2)^{\nu_1/2}[\Gamma(\nu_1/2)]^{-1} .$$

By comparing the likelihood function given by (2.22) and the prior given by (2.24), the conjugate relationship between the likelihood function and the prior density function is evident.

In most of the analysis, the new parameters introduced in the prior distribution given by (2.24), namely p_1, v_1, and ν_1, will have obvious meaning. p_1 corresponds to the number of parameters in $\underline{\theta}$, namely 3, v_1 is the prior mean of σ^2, and ν_1 is the degree of belief in the prior mean of σ^2. In the cases which call for vague or noninformative priors, these parameters take on special values. Bayesian analysis with vague priors is considered in Chapter 4.

2.3 Posterior Distribution

In Bayesian analysis the posterior distribution on the unknown parameters is proportional to the prior distribution of the parameters multiplied by the likelihood function. As in the discussion on the prior distributions, the situations of known and unknown precision are considered separately, with the case of known precision considered first.

2.3.1 Posterior Distribution, Precision Known

Let $f_2(\underline{\theta}/h, W(\gamma), \underline{Y})$ be the posterior density for $\underline{\theta}$, given h, $W(\gamma)$, and $\underline{Y}$. The notation $W(\gamma)$ is used to emphasize the dependence of the posterior distribution on the value of the join point. By Bayes' formula this density is proportional to

$$\ell(\underline{\theta}/h, W(\gamma), Y) \cdot f_1(\underline{\theta}/\gamma, h)$$

where $f_1(\underline{\theta}/\gamma, h)$ is, by (2.21), proportional to

$$\exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu}_1)'\Sigma_1^{-1}(\underline{\theta}-\underline{\mu}_1)\}\ , \tag{2.25}$$

and $\ell(\underline{\theta}/h, W(\gamma), Y)$ is, by (2.10) proportional to

$$\exp\{-\frac{h}{2}[(\underline{Y}-W\underline{\hat{\theta}})'(\underline{Y}-W\underline{\hat{\theta}}) + (\underline{\theta}-\underline{\hat{\theta}})'W'W(\underline{\theta}-\underline{\hat{\theta}})]\}\ .$$

But, since h is assumed known and $(Y-W\hat{\theta})'(Y-W\underline{\hat{\theta}})$ does not

involve the vector $\underline{\theta}$, for the case when precision is known the only part of the likelihood function which we need to consider in determining the posterior distribution is

$$\exp\{-\frac{h}{2}(\underline{\theta}-\hat{\underline{\theta}})'W'W(\underline{\theta}-\hat{\underline{\theta}})\} . \tag{2.26}$$

Multiplying (2.25) and (2.26) yields the result that the posterior density function for $\underline{\theta}$ is proportional to

$$\exp\{-\frac{h}{2}[(\underline{\theta}-\underline{\mu_1})'\Sigma_1^{-1}(\underline{\theta}-\underline{\mu_1})+(\underline{\theta}-\hat{\underline{\theta}})'W'W(\underline{\theta}-\hat{\underline{\theta}})]\} .$$

Consider the terms multiplied by $(-h/2)$. By rearranging these terms we get

$$\begin{aligned}(\underline{\theta}-\underline{\mu_1})'\Sigma_1^{-1}(\underline{\theta}-\underline{\mu_1}) + (\underline{\theta}-\hat{\underline{\theta}})'W'W(\underline{\theta}-\hat{\underline{\theta}}) \\ = \underline{\theta}'(\Sigma_1^{-1}+W'W)\underline{\theta} - 2(\underline{\mu_1}'\Sigma_1^{-1}+\hat{\underline{\theta}}'W'W) + \underline{\mu_1}'\Sigma_1^{-1}\underline{\mu_1}+\hat{\underline{\theta}}'W'W\hat{\underline{\theta}} .\end{aligned} \tag{2.27}$$

This is possible since $\underline{\mu_1}'\Sigma_1^{-1}\underline{\theta}$ and $\hat{\underline{\theta}}'W'W\underline{\theta}$ are both scalars, thus giving

$$\underline{\mu_1}'\Sigma_1^{-1}\underline{\theta} = (\underline{\mu_1}'\Sigma_1^{-1}\underline{\theta})' = (\underline{\theta}'\Sigma_1^{-1}\underline{\mu_1}) ,$$

and

$$\hat{\underline{\theta}}'W'W\underline{\theta} = (\hat{\underline{\theta}}'W'W\underline{\theta})' = \underline{\theta}'W'W\hat{\underline{\theta}}' .$$

Since the prior distribution is conjugate for the likelihood, the posterior density must be of the same functional form as the prior. Using this property, we know that $f_2(\underline{\theta}/h, W(\gamma), \underline{Y})$ must be a trivariate normal density function with mean vector $\underline{\mu_2}$ and dispersion matrix $h^{-1}\Sigma_2$. With this notation the exponent of the density function of the trivariate normal, ignoring the multiplier $(-h/2)$, is

$$\underline{\theta}'\Sigma_2^{-1}\underline{\theta} - 2\underline{\mu_2}\Sigma_2^{-1}\underline{\theta} + \underline{\mu_2}'\Sigma_2^{-1}\underline{\mu_2} . \tag{2.28}$$

But the terms in (2.27) and (2.28) are both the exponent, ignoring the multiplier, of the same trivariate normal density function. Therefore, using the equality of these two sets of terms, we can determine the values for $\underline{\mu_2}$ and Σ_2 as functions of the prior distribution and the likelihood function. From these two equations it follows that

$$\underline{\theta}'(\Sigma_1^{-1}+W'W)\underline{\theta} = \underline{\theta}'(\Sigma_2^{-1})\underline{\theta}, \quad \text{for all } \underline{\theta} .$$

In order for this to be true, it is necessary that

$$\Sigma_2^{-1} = \Sigma_1^{-1} + W'W , \tag{2.29}$$

or

$$\Sigma_2 = (\Sigma_1^{-1}+W'W)^{-1} .$$

Also, the two different forms have the correspondence that

$$\underline{\mu}_2'\Sigma_2^{-1}\underline{\theta} = (\underline{\mu}_1'\Sigma_1^{-1}+\hat{\underline{\theta}}'W'W)\underline{\theta}, \quad \text{for all} \quad \underline{\theta} .$$

This yields

$$\underline{\mu}_2'\Sigma_2^{-1} = \underline{\mu}_1'\Sigma_1^{-1} + \hat{\underline{\theta}}'W'W ,$$

or

$$\underline{\mu}_2' = (\underline{\mu}_1'\Sigma_1^{-1}+\hat{\underline{\theta}}'W'W)\Sigma_2$$

or

$$\underline{\mu}_2 = \Sigma_2(\Sigma_1^{-1}\underline{\mu}_1+W'W\hat{\underline{\theta}}) . \tag{2.30}$$

Therefore, the posterior distribution of $\underline{\theta}$ for known h, is the trivariate normal with mean given by (2.30) and dispersion matrix

$$h^{-1}\Sigma_2 = [h(\Sigma_1^{-1}+W'W)]^{-1} .$$

The posterior parameter of most concern in the analysis when the join point and the precision parameter are assumed known is the mean vector $\underline{\mu}_2$. Equation (2.30) tells us that this parameter vector is a weighted combination of the prior mean vector $\underline{\mu}_1$ and the vector of least squares estimates $\hat{\underline{\theta}}$, with the weighting factors being $\Sigma_2\Sigma_1^{-1}$ and $\Sigma_2W'W$, respectively.

2.3.2 Posterior Distribution, Precision Unknown

In the situation where the precision parameter h is unknown, the posterior distribution of $\underline{\theta}$ becomes somewhat more complicated.

Bayes' theorem is used in the same way, but the prior distribution and the likelihood functions are different from those used in the previous section. The joint posterior density function $g_2(\underline{\theta}, h/W(\gamma), \underline{Y})$ is proportional to

$$g_1(\underline{\theta}, h)\ell(\underline{\theta}, H/W(\gamma), Y)$$

where $g_1(\underline{\theta}, h)$, given by (2.24) is proportional to

$$h^{p_1/2}|\Sigma_1|^{-1/2}\exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu_1})'\Sigma_1^{-1}(\underline{\theta}-\underline{\mu_1})\}$$
$$\times h^{\nu_1/2-1}\exp\{-h\nu_1 v_1/2\}(\nu_1 v_1/2)^{\nu_1/2}[\Gamma(\nu_1/2)]^{-1},$$

and $\ell(\underline{\theta}, h/W(\gamma), \underline{Y})$, given by (2.22), is proportional to

$$h^{n/2}\exp\{-\frac{h}{2}(\underline{\theta}-\underline{\hat{\theta}})'W'W(\underline{\theta}-\underline{\hat{\theta}})\}\exp\{(-\frac{h}{2})(n-3)s^2\}.$$

Now, since we are still working with a conjugate prior density function, we know that the posterior density function must be of the same form as the prior density function. Because of this, we know that the posterior density $g_2(\underline{\theta}, h/W(\gamma), \underline{Y})$ is proportional to

$$h^{p_2/2}|\Sigma_2|^{-1/2}\exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu_2})'\Sigma_2^{-1}(\underline{\theta}-\underline{\mu_2})\} \tag{2.31}$$
$$\times h^{\nu_2/2-1}\exp\{-h\nu_2 v_2/2\},$$

where p_2, Σ_2, $\underline{\mu}_2$, ν_2, and v_2 are the parameters of the posterior distribution which correspond, respectively, to p_1, Σ_1, $\underline{\mu}_1$, ν_1 and v_1, the parameters of the prior density function. Using a method similar to that used in the previous section, the relationships of the posterior parameters to the prior parameters and the statistics of the likelihood function can be determined.

As in the case where the precision parameter h is known, we have

$$\Sigma_2 = (\Sigma_1^{-1} + W'W)^{-1}, \tag{2.32}$$

and

$$\underline{\mu}_2 = \Sigma_2(\Sigma_1^{-1}\underline{\mu}_1 + W'W\hat{\underline{\theta}}) .$$

For the three parameters which did not exist in the case where the precision was assumed known, the following relations hold:

$$p_2 = r(\Sigma_2^{-1}) = 3,$$

$$\nu_2 = n + \nu_1 + p_1 - p_2 = n + \nu_1 + p_1 - 3 = n + \nu_1 ;$$

and

$$v_2 = \frac{1}{\nu_2}\{\nu_1 v_1 + \underline{\mu}_1'\Sigma_1^{-1}\underline{\mu}_1 + (n-3)s^2 + \hat{\underline{\theta}}'W'W\hat{\underline{\theta}} - \underline{\mu}_2'\Sigma_2^{-1}\underline{\mu}_2\} . \tag{2.33}$$

In using the posterior density function for $\underline{\theta}$ and h, many different aspects can be considered. The posterior density function

given by (2.31) with parameters given by (2.32) and (2.33) contains all of the present knowledge of the unknown parameters after the prior belief has been updated by the data resulting from the experiment. For any given value of h, the posterior density function of $\underline{\theta}$ is a trivariate normal. Because of this fact the posterior distribution of any subset or linear combination of the vector $\underline{\theta}$ can be easily obtained and analyzed using multivariate normal theory. The marginal posterior distribution of h is of the same form as (2.23), and thus also belongs to the gamma family. In a particular problem the experimenter might want to consider some or all of the possible conditional and marginal distributions. At this time we will be content to merely look at the posterior parameters. In Chapter 4 the various conditional and marginal distributions will be studied when the join point is unknown.

Throughout the above sections on posterior density functions, Σ_2^{-1} appears many times. Σ_2^{-1} is nonsingular, and thus we are working with a non-degenerate trivariate normal distribution as our posterior. In both sections on posterior distributions, the parameter Σ_2^{-1} is defined as

$$\Sigma_2^{-1} = \Sigma_1^{-1} + W'W .$$

Both Σ_1^{-1} and $W'W$ are non-negative definite matrices.

Therefore, if either of these two matrices is positive definite, it follows that Σ_2^{-1} is positive definite and Σ_2 is positive definite. Now, in most cases both of the matrices in question, namely Σ_1^{-1} and $W'W$, are non-singular. Suppose the rank of Σ_1^{-1} is less than 3 (full rank), then the prior distribution is saying that one of the three unknown parameters in our reparameterized regression model, namely θ_1, θ_2, and θ_3, is a linear combination of the other two, or that one of the parameters is known. That is, information on two of the unknown parameters is all that would be necessary in the problem. If this situation arises, the analysis of the problem changes accordingly. The likelihood function in this situation consists of two unknown parameters (not including the variance, or precision), and is expressed in an appropriate form. For this case of a prior having a bivariate normal distribution and a likelihood function with a design matrix of rank two, the analysis proceeds similar to that developed for the three parameter case. Since this situation is actually a separate problem, in this paper we assume that the dispersion matrix of the prior distribution is of rank three.

Consideration of the matrix $W'W$ also yields the result that in all practical situations this matrix is nonsingular. Thus, we will assume that Σ_2^{-1} is always nonsingular, and that the posterior distribution consists, in whole or part of a non-degenerate trivariate normal distribution.

In this chapter we studied the expression of the likelihood function, the development of the prior distribution, and the parameters of the resulting posterior distribution when the join point of the two linear regression segments is known. In addition to knowledge of the join point, the experimenter must specify values for the following parameters; $\underline{\mu}_1$, Σ_1, p_1, ν_1, and v_1. When specifying values for $\underline{\mu}_1$ and Σ_1, the values can be specified for $\underline{\mu}_1(\phi)$ and $\Sigma_1(\phi)$, or, if the experimenter has much knowledge of the dependencies that should exist within the parameters of the vector $\underline{\theta}$ he can choose $\underline{\mu}_1$ and Σ_1 directly. After choosing these parameters and assuming the conjugate priors used here, the posterior density function is given by (2.31) with parameters defined by (2.32) and (2.33). This assumes h is unknown. In order to better illustrate how the Bayesian approach for known join point works, a detailed numerical example will now be considered.

2.4 Example

The data in this example comes from Hudson (10). We assume that the join point occurs at $x = 4$. The data points are

X	1	2	3	4	5	6
Y	1	2	4	7	3	1

The data points are plotted in Figure 5. We will now consider separately each of the three parts of the Bayesian analysis for this data: (1), the likelihood function; (2), the prior distribution; and (3), the posterior distribution. Since the forms of the prior and posterior distributions have been determined in the previous sections, we will concern ourselves primarily with the prior and posterior parameters.

(1) The likelihood function. The two-phase linear model for the data in matrix notation is

$$\underline{Y} = W\underline{\theta} + \underline{\epsilon} ,$$

where

$$\underline{Y} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 7 \\ 3 \\ 1 \end{pmatrix}, \quad \underline{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}, \quad \underline{\epsilon} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{pmatrix},$$

and

$$W = XA(4) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 0 & 1 & 5 \\ 0 & 0 & 1 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & -4 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 4 & 0 \\ 1 & 5 & 1 \\ 1 & 6 & 2 \end{bmatrix} .$$

If we assume that the errors are normally distributed, only three terms from the likelihood function play a part in determining the

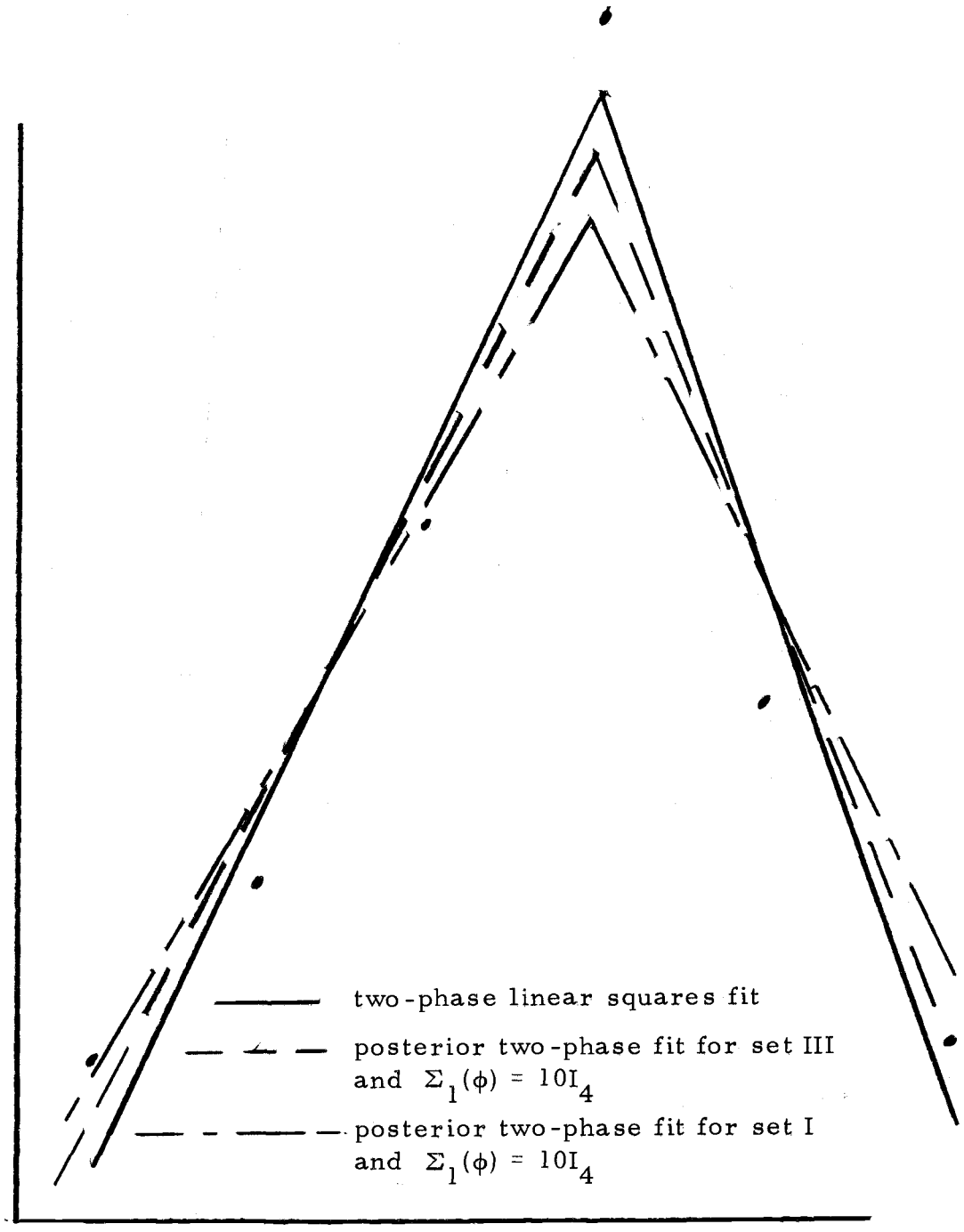

Figure 5. Plot of data points for example 2.4.

posterior parameters. These three terms are $\hat{\underline{\theta}}$, W'W, and $s^2 = (n-3)^{-1}(\underline{Y}-W\hat{\underline{\theta}})'(\underline{Y}-W\hat{\underline{\theta}})$. In this problem these terms are

$$(2.34) \qquad \hat{\underline{\theta}} = (W'W)^{-1}W'\underline{Y}, \qquad s^2 = (6-3)^{-1}(Y-W\hat{\underline{\theta}})'(Y-W\hat{\underline{\theta}})$$

$$\hat{\underline{\theta}} = \begin{pmatrix} -1.37 \\ 1.92 \\ -4.71 \end{pmatrix}, \qquad s^2 = (3)^{-1}(1.39) = .463$$

and

$$W'W = \begin{bmatrix} 6 & 21 & 3 \\ & 91 & 17 \\ & & 5 \end{bmatrix}.$$

2. The prior distribution. Since this is an artificial problem taken from an outside source, there is no method available to determine a legitimate prior distribution. For this reason a variety of different values for the prior parameters will be used.

As mentioned previously, it is necessary to determine either $\underline{\mu}_1$ and Σ_1 or $\underline{\mu}_1(\phi)$ and $\Sigma_1(\phi)$. Because of the nature of the problem, we will start with $\underline{\mu}_1(\phi)$ and $\Sigma_1(\phi)$. In (2.20) the expressions for $\underline{\mu}_1$ and Σ_1 are given for any choice of $\underline{\mu}_1(\phi)$ and $\Sigma_1(\phi)$. In these expressions the only dependency on the join point is contained in the vector $\underline{\delta}'(\gamma)$. In this example,

$$\underline{\delta}'(\gamma) = \underline{\delta}'(4) = (1 \quad 4 \quad -1 \quad -4).$$

Let us first look at the dispersion matrix, $\Sigma_1(\phi)$. In this example we will assume the unconstrained prior dispersion matrix

takes the form

$$\Sigma(\phi) = (10)^k I_4, \quad \text{where} \quad k = -1, 0, 1, 2, 3.$$

By varying k, we indicate whether the belief in the prior parameter $\underline{\mu}_1(\phi)$ is strong or weak. For small values of k, the prior belief is very strong. In preliminary studies on this problem, values of k which were less than (-1) gave posterior means nearly identical to prior means. As k gets larger, the effects of the prior belief lessens. For k = 3, the effect is negligible.

If $\Sigma_1(\phi) = (10)^k I_4$, then be evaluating (2.20),

$$\Sigma_1 = (10)^k \begin{bmatrix} .9705 & -.1176 & .2353 \\ & .5294 & -.0588 \\ & & .1176 \end{bmatrix}. \tag{2.35}$$

The above dispersion matrix has some characteristics which should be noted. The parameter θ_3 has the smallest prior variance of the three unknown parameters, and θ_2 is negatively correlated with the other two parameters.

Let us now consider the prior mean vector $\underline{\mu}_1$. For an arbitrary prior mean vector $\underline{\mu}_1'(\phi) = (\mu_1\ \mu_2\ \mu_3\ \mu_4)$ with $\Sigma_1(\phi) = (10)^k I_4$, (2.20) gives us

$$(2.36) \qquad \underline{\mu}_1 = \frac{1}{2(1+\gamma^2)} \begin{bmatrix} \mu_1 + \mu_3 + \gamma(\mu_4-\mu_2) + 2\gamma^2\mu_1 \\ 2\mu_2 + \gamma(\mu_3-\mu_1) + \gamma^2(\mu_2+\mu_4) \\ 2\gamma(\mu_1-\mu_3) + 2(\mu_4-\mu_2) \end{bmatrix},$$

or where $\gamma = 4$,

$$\underline{\mu}_1 = \frac{1}{34} \begin{bmatrix} \mu_1 + \mu_3 + 4(\mu_4-\mu_2) + 32\mu_1 \\ 2\mu_2 + 4(\mu_3-\mu_1) + 16(\mu_2+\mu_4) \\ 8(\mu_1-\mu_3) + 2(\mu_4-\mu_2) \end{bmatrix}.$$

As should be expected, the different values of k have no effect on $\underline{\mu}_1$. However, if we were to change the form of $\Sigma_1(\phi)$, the expression above for $\underline{\mu}_1$ would not be the same.

The three sets of values for $\underline{\mu}_1(\phi)$ which we consider are

$$\text{(I)} \quad \underline{\mu}_1(\phi) = \begin{pmatrix} 0 \\ 1 \\ 13 \\ -2 \end{pmatrix}, \quad \text{(II)} \quad \underline{\mu}_1(\phi) = \begin{pmatrix} -1 \\ 2 \\ 10 \\ -2 \end{pmatrix}, \quad \text{(III)} \quad \underline{\mu}_1(\phi) = \begin{pmatrix} -1 \\ 2 \\ 15 \\ -2 \end{pmatrix}.$$

The first and second set of values were "rough guesses" made after looking at a plot of the data. The third set was chosen after observing the estimates derived by the least squares technique. These three sets of parameters yielded the following values for $\underline{\mu}_1$, which will be denoted in any following discussion and in the tables as set I, set II, and set III.

(2.37)
$$\underline{\mu}_1 = \begin{matrix} \text{I} \\ \begin{pmatrix} .029 \\ 1.12 \\ -3.24 \end{pmatrix} \end{matrix}, \quad \underline{\mu}_1 = \begin{matrix} \text{II} \\ \begin{pmatrix} -1.15 \\ 1.41 \\ -2.82 \end{pmatrix} \end{matrix}, \quad \underline{\mu}_1 = \begin{matrix} \text{III} \\ \begin{pmatrix} -1.0 \\ 2.0 \\ -4.0 \end{pmatrix} \end{matrix}.$$

The other prior parameters are p_1, v_1, and ν_1. In this problem $p_1 = 3$. For v_1 and ν_1 we will consider various possible values when we look at the posterior parameters.

(3) Posterior distribution. In the posterior distribution when h is unknown, there are five parameters to consider. These are $\underline{\mu}_2$, Σ_2, p_2, v_2, and ν_2. Two of these parameters are easily obtained. They are $p_2 = 3$, and $\nu_2 = 6 + \nu_1$.

Let us first consider the posterior dispersion matrix, Σ_2. Since we are considering five different values for k, there are five different expressions for the posterior dispersion matrix. These expressions do not depend on $\underline{\mu}_1$, so they hold true for the three sets of values of $\underline{\mu}_1$. The results are found in Table 2.1. The matrix $(W'W)^{-1}$ is also presented to show the convergence of Σ_2 to $(W'W)^{-1}$ as k gets large. The determinants of these matrices are also presented. The terms in the posterior dispersion matrix are closest to zero for the cases where k is smallest. As k increases and the degree of belief in the prior decreases, the terms of the posterior matrix become larger in absolute value until the case where $k = 3$, at which point the posterior dispersion matrix is

nearly identical to $(W'W)^{-1}$, the dispersion matrix of the least squares estimates. The determinant of the posterior dispersion matrix is always less than or equal to the determinant of $(W'W)^{-1}$.

Table 2.1. Values of the posterior dispersion matrix Σ_2 when $\Sigma_1(\phi) = (10)^k I_4$.

				$\lvert\Sigma_2\rvert$
	.0869	-.0217	.0217	
k = -1		.0148	-.0069	.0000046
			.0137	
	.5339	-.1488	.1460	
k = 0		.0542	-.0512	.0003
			.0939	
	1.173	-.3543	.4490	
k = 1		.1280	-.1893	.0037
			.4580	
	1.413	-.4539	.6845	
k = 2		.1743	-.3137	.0078
			.8337	
	1.452	-.4716	.7311	
k = 3		.1831	-.3390	.0086
			.9115	
	1.456	-.4736	.7368	$\lvert(W'W)^{-1}\rvert$
$(W'W)^{-1}$		.1842	-.3421	= .0087
			.9210	

Probably the most important posterior parameter in our example is the posterior mean vector for $\underline{\theta}$, namely $\underline{\mu}_2$. In addition to the sufficient statistics of the likelihood function, by (2.32) it depends on $\underline{\mu}_1$, Σ_1^{-1}, and Σ_2. The results for the various sets of $\underline{\mu}_1$ and the five possible sets of Σ_1 are contained in Table 2.2. For the

Table 2.2. Posterior values of $\underline{\mu}_2$.

Prior Set for $\underline{\mu}_1$	Param.	Prior Mean	$\Sigma(\phi) = (10)^k I_4$				
			k = -1	k = 0	k = 1	k = 2	k = 3
I	θ_1	.029	.023	-.278	-.889	-1.28	-1.36
	θ_2	1.12	1.29	1.42	1.68	1.88	1.92
	θ_3	-3.24	-3.25	-3.40	-3.99	-4.58	-4.70
II	θ_1	-1.15	-1.06	-.810	-.854	-1.26	-1.36
	θ_2	1.41	1.48	1.46	1.60	1.86	1.91
	θ_3	-2.82	-2.82	-2.91	-3.64	-4.50	-4.68
III	θ_1	-1.00	-1.06	-1.06	-1.16	-1.33	-1.36
	θ_2	2.00	1.77	1.73	1.80	1.90	1.92
	θ_3	-4.00	-4.01	-4.06	-4.33	-4.64	-4.70

case where $k = 3$, the posterior means are very close for the three different sets of $\underline{\mu}_1$. They are also very close to the least squares estimates given by (2.34). As would be expected, the posterior means do not change much where set III is the prior mean for $\underline{\mu}_1$. This prior was chosen after observing the least squares estimates. Reading across the table for any of the three choices of $\underline{\mu}_1$ as k goes from (-1) to 3, the value of the posterior mean for each parameter changes from a value very close to the prior mean to a value very close to the least squares estimate of that parameter. In some ways, this example answers the argument against being able to give adequate priors. By letting the prior dispersion matrix get large, the effect of the prior mean becomes negligible. However, the prior distribution does let you enter any significant prior knowledge into the model.

The last posterior parameter to consider is v_2, which is a posterior estimate of σ^2. By (2.33) it depends on the prior mean vector, the prior dispersion matrix, the prior parameter v_1, and the prior parameter ν_1. It also depends on $\hat{\underline{\theta}}$, $W'W$, $\underline{\mu}_2$, and Σ_2^{-1}. The least squares estimate of σ^2 is .463. In our study we looked at the posterior parameter v_2 when v_1 takes on the values .25, .5, and 1.0. With each of these three values, we considered five possible values for ν_1, the degree of belief in v_1; namely 1, 3, 6, 12, and 30. The values 1 and 3 indicate weak belief in the prior, the value 6 is giving the prior and likelihood equal weight,

and 12 and 30 are values where the prior belief is much stronger than the likelihood. As mentioned above, v_2 depends on $\underline{\mu}_1$ and Σ_1, thus yielding 15 values for each combination of $\underline{\mu}_1$ and Σ_1. Since the only effect of changing Σ_1 is a decreasing of the values as k increases, we will only show the cases where $\Sigma_1(\phi) = I_4$, and $\Sigma_1(\phi) = 100\ I_4$. The results are presented in Table 2.3. Values of ν_2 are also included.

By studying the results in the table, it is obvious that the posterior distribution corresponding to set III as $\underline{\mu}_1$ has the smallest posterior values for v_2. In Chapter 3 we will show that the variance of any posterior distribution on $\underline{\theta}$, whether marginal or conditional, depends only on ν_2, v_2, and Σ_2. In any situation where ν_1 (and thus ν_2), v_1, and Σ_2 are chosen, the smallest of the three values for v_2 comes from Table 2.3c.

Therefore, if we were to compare our three choices for $\underline{\mu}_1$, the posterior distribution corresponding to set III would be preferred.

Much time has been spent on this example, and much time will be spent on the example in the following chapter. It is felt that only through these detailed examples can many of the ramifications of this Bayesian approach be seen. Some criticism might be made of the fact that so many possible values of the various prior parameters were used. This approach was taken to illustrate how change in the various prior parameters changes the posterior parameters. The Bayesian

analysis of a realistic problem might not proceed in this manner, but it is hoped that the example illustrates the multitude of results that can come from one set of data, depending on prior knowledge of the situation.

Table 2.3. Posterior values for v_2.

v_2	v_1	$\Sigma_1(\phi) = I_4$ .25	$\Sigma_1(\phi) = I_4$ .50	$\Sigma_1(\phi) = I_4$ 1.0	$\Sigma_1(\phi) = 100 I_4$ .25	$\Sigma_1(\phi) = 100 I_4$ .50	$\Sigma_1(\phi) = 100 I_4$ 1.0
a) Posterior values for v_2 for $\underline{\mu}_1$ = set I.							
7	1	.549	.585	.656	.265	.301	.372
9	3	.482	.566	.732	.262	.345	.512
12	6	.424	.549	.799	.259	.384	.634
18	12	.366	.533	.866	.256	.422	.756
36	30	.308	.516	.933	.253	.461	.878
b) Posterior values for v_2 for $\underline{\mu}_1$ = set II							
7	1	.950	.986	1.06	.307	.343	.414
9	3	.794	.878	1.04	.294	.378	.544
12	6	.658	.783	1.03	.283	.408	.658
18	12	.522	.689	1.02	.272	.439	.772
36	30	.386	.594	1.01	.261	.469	.886
c) Posterior values for v_2 for $\underline{\mu}_1$ = set III							
7	1	.339	.374	.446	.244	.280	.351
9	3	.319	.402	.569	.245	.328	.495
12	6	.302	.427	.677	.246	.371	.621
18	12	.284	.451	.784	.248	.414	.748
36	30	.267	.476	.892	.249	.457	.874

3. A BAYESIAN PROCEDURE FOR THE ANALYSIS OF A TWO-PHASE REGRESSION MODEL: II. JOIN POINT UNKNOWN

The analysis in the previous chapter assumed that the join point was known. However, in most situations, the join point is not known, and one of the principal objectives of the analysis is to make inference on the unknown value of the join point. The format of this chapter is similar to that of the previous chapter. In most of the following discussion we assume that the precision is unknown. The case where the precision is known is merely a special case of the more general situation, and will be only briefly considered.

3.1 The Likelihood Function

Returning to the original model of the two-phase linear regression problem, we have $E(\underline{Y}) = X\underline{\phi}$, where

$$X = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ 1 & x_2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_r & 0 & 0 \\ 0 & 0 & 1 & x_{r+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & x_n \end{bmatrix}, \quad \text{and} \quad \underline{\phi} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \beta_0 \\ \beta_1 \end{bmatrix},$$

if we assume that $x_r \leq \gamma < x_{r+1}$. As is obvious from the form of the

above design matrix X, in order to explicitly write out the model, we must know between which two x-observations the join point occurs. Because of this, for the situation where the join point is not known we will use conditional likelihood functions; that is, we will have k likelihood functions where each function is conditioned on the join point γ being equal to a certain value, c_i, $i = 1, \ldots, k$. In doing this we are assuming that γ can occur at one of only a finite number of different values of the independent variable x. The conditional likelihood functions are designated by $\ell_i(\underline{\theta}, h/W(c_i), \underline{Y})$, $i = 1, \ldots, k$, where $W(c_i) = XA(c_i)$.

$$(3.1) \qquad W(c_i) = \begin{bmatrix} 1 & x_1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_i & 0 \\ 1 & x_{i+1} & x_{i+1}-c_i \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n-c_i \end{bmatrix}, \quad \text{if} \quad c_i \in [x_i, x_{i+1}).$$

The possible values at which the join point can occur in a given problem can be uncountable. However, if the prior distribution for γ is continuous, problems result in formulating the likelihood function as a function of the sufficient statistics, because W is a function of the exact value of the join point and the sufficient statistics depend on W. The method of analysis depends on the information of the sample

being represented by sufficient statistics, so that this information can be combined with the parameters of the conjugate prior distribution. In order to alleviate this problem, the possible values of the join point are "discretized," giving us only a finite number of different join points. Using this approach some information is lost, but this loss can be minimized by suitable choices of the values c_i, $i = 1, \ldots, k$, where k is the number of different values at which we are assuming the join point can occur. None of the following analysis depends on any assumptions on the possible values of the join point, other than the restriction that there are only a finite number of different values. More than one value of c may be placed in a given interval, and every possible interval need not contain a possible value of a join point. By placing more than one value in the intervals which are most likely to contain the join point, the amount of information lost by "discretizing" is reduced.

Except in situations where confusion might result we will use the notation W_i in place of $W(c_i)$. The least squares estimates of the vector $\underline{\theta}$ for $\gamma = c_i$, is $\underline{\hat{\theta}}_i$, where

$$\underline{\hat{\theta}}_i = (W_i'W_i)^{-1}W_i'Y .$$

Using this notation the conditional likelihood for $\gamma = c_i$, namely $\ell_i(\underline{\theta}, h/W(c_i), \underline{Y})$, is proportional to

$$(3.2) \qquad h^{n/2} \exp\{-\frac{h}{2}(\underline{Y}-W_i\underline{\hat{\theta}})'(\underline{Y}-W_i\underline{\hat{\theta}}) - \frac{h}{2}(\underline{\theta}-\underline{\hat{\theta}}_i)'W_i'W_i(\underline{\theta}-\underline{\hat{\theta}}_i)\} .$$

Sometimes for convenience of notation, we will use s_i^2, where

$$s_i^2 = (n-2)^{-1}(\underline{Y}-W_i\underline{\hat{\theta}}_i)'(\underline{Y}-W_i\underline{\hat{\theta}}_i) .$$

If this is done, the likelihood function is proportional to

$$(3.3) \qquad h^{3/2} \exp\{-\frac{h}{2}[(n-3)s_i^2 + (\underline{\theta}-\underline{\hat{\theta}}_i)'W_i'W_i(\underline{\theta}-\underline{\hat{\theta}}_i)]\} .$$

3.2 Joint Prior Distribution

We assume that the abscissa of the join point, namely γ, can occur at only a finite number of points on the real line. For this reason the prior distribution for γ takes the form of a discrete probability function.

Let the marginal prior belief that the join point γ occurs at the value c_i be designated by

$$P(\gamma=c_i), \quad i = 1, \ldots, k.$$

For any particular value of γ, say c_i, the joint prior belief on all the unknown parameters is the product of the prior belief that the join point equals c_i and the conditional prior distribution on the other parameters, given that the join point occurs at c_i. Thus, the

joint prior belief for $\underline{\theta}$, h, $\gamma = c_i$, is

$$g_{1i}(\underline{\theta}, h/\gamma=c_i)P(\gamma=c_i), \quad i = 1, \ldots, k.$$

We assume that $g_{1i}(\underline{\theta}, h/\gamma=c_i)$ has the same functional form as $g_1(\underline{\theta}, h)$ given by (2.24). Whereas $g_1(\underline{\theta}, h)$ is the prior distribution when γ is known, $g_{1i}(\underline{\theta}, h/\gamma=c_i)$ is the prior distribution when we assume $\gamma = c_i$. The only difference in the two functions is that the value of c_i now appears wherever γ previously appeared. This holds true for the k different functions, $g_{1i}(\underline{\theta}, h/\gamma=c_i)$. Therefore, $g_{1i}(\underline{\theta}, h/\gamma=c_i)$ is proportional to

$$(3.4) \qquad |\Sigma_{1i}|^{-1/2} h^{p_{1i}/2} \exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu}_{1i})'\Sigma_{1i}^{-1}(\underline{\theta}-\underline{\mu}_{1i})\}$$

$$\times (\nu_{1i}v_{1i}/2)^{\nu_{1i}/2} h^{\nu_{1i}/2-1} \exp\{-h\nu_{1i}v_{1i}/2\}[\Gamma(\nu_{1i}/2)]^{-1}.$$

In the above function the only parameters that are explicitly dependent on the value of c_i are $\underline{\mu}_{1i}$ and Σ_{1i}^{-1}. In order that no generality is lost, assume that p_{1i}, ν_{1i}, and v_{1i} may take on different values for each of the possible values of the join point. This formulation imposes no particular restrictions on the values that these three parameters may assume, except that they must all be non-negative. The formulation does impose restrictions on the parameters $\underline{\mu}_{1i}$ and Σ_{1i}. By the method used in Chapter 2 to find

$\underline{\mu}_1$ and Σ_1 for a known join point, it is easy to show that

$$\underline{\mu}_{1i} = B(I-H(c_i))\underline{\mu}_1(\phi) , \tag{3.5}$$

and

$$\Sigma_{1i} = B(I-H(c_i))\Sigma_1(\phi)B' ,$$

where

$$H(c_i) = \Sigma_1(\phi)\underline{\delta}(c_i)[\underline{\delta}'(c_i)\Sigma_1(\phi)\underline{\delta}(c_i)]^{-1}\underline{\delta}'(c_i)$$

$$\underline{\delta}'(c_i) = (1 \quad c_i \quad -1 \quad -c_i) ,$$

and B is given by (2.18). Thus, the prior knowledge, or belief, for the unknown join point and the vector $\underline{\theta}$ and parameter h is contained in the marginal probabilities $P(\gamma=c_i)$, and the k sets of parameters $(p_{1i}, \nu_{1i}, v_{1i}, \underline{\mu}_{1i}, \Sigma_{1i})$, $i = 1, \ldots, k$, where the values of some of the parameters might be constant for different values of the c_i's.

3.3 Joint Posterior Distribution

Since we are assuming γ has a discrete distribution, the joint posterior distribution will consist of k parts, one for each possible value of γ for which we give a prior probability. Denote the ith part of the posterior distribution as

$$g_{2i}(\underline{\theta}, h/W(c_i), \underline{Y})P(\gamma=c_i/X, Y), \quad i = 1, \ldots, k,$$

where $g_{2i}(\underline{\theta}, h/W(c_i), \underline{Y})$ is proportional to

$$\ell_i(\underline{\theta}, h/W(c_i), \underline{Y}) g_{1i}(\underline{\theta}, h/\gamma=c_i) .$$

We can also write

(3.6)
$$P(\gamma=c_i/X, Y) g_{2i}(\underline{\theta}, h/W(c_i), Y)$$

$$= \frac{\ell_i(\underline{\theta}, h/W(c_i), Y) g_{1i}(\underline{\theta}, h/\gamma=c_i) P(\gamma=c_i)}{T'}$$

where

$$T' = \sum_{i=1}^{k} \int_H \int_\Theta \ell_i(\underline{\theta}, h/W(c_i), Y) g_{1i}(\underline{\theta}, h/\gamma=c_i) P(\gamma=c_i) \underline{d\theta} dh$$

and H and Θ denote the domain of positive probability for h and $\underline{\theta}$, respectively.

In order to more explicitly state the posterior density functions for h and $\underline{\theta}$, and the posterior probability function for γ, we need to determine the value of T'. From previous developments we know that the numerator of (3.6) is proportional to

$$h^{n/2} \exp\{-\frac{h}{2}[(n-3)s_i^2 + (\underline{\theta}-\underline{\hat{\theta}}_i)' W_i' W' (\underline{\theta}-\underline{\hat{\theta}}_i)]\}$$

$$\times |\Sigma_{1i}|^{-1/2} h^{p_1/2} \exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu}_{1i})' \Sigma_{1i}^{-1} (\underline{\theta}-\underline{\mu}_{1i})\}$$

$$\times h^{\nu_{1i}/2-1} \exp\{-h\nu_{1i}v_{1i}/2\} (\nu_{1i}v_{1i}/2)^{\nu_{1i}/2} [\Gamma(\nu_{1i}/2)]^{-1} P(\gamma=c_i) .$$

The above function can be arranged so that it equals

$$(3.7)\quad |\Sigma_{1i}|^{-1/2} h^{(n+p_{1i})/2} \exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu}_{2i})'\Sigma_{2i}^{-1}(\underline{\theta}-\underline{\mu}_{2i})\}$$

$$\times h^{\nu_{1i}/2-1} \exp\{-h\nu_{2i}v_{2i}/2\}(\nu_{1i}v_{1i}/2)^{\nu_{1i}/2}[\Gamma(\nu_{1i}/2)]^{-1}P(\gamma=c_i)$$

where

$$(3.8)\quad \underline{\mu}_{2i} = \Sigma_{2i}(\Sigma_{1i}^{-1}\underline{\mu}_{1i}+W_i'W_i\hat{\underline{\theta}}_i) ,$$

$$\Sigma_{2i}^{-1} = \Sigma_{1i}^{-1} + W_i'W_i ,$$

$$\nu_{2i} = n + \nu_{1i} + p_{1i} - p_{2i}$$

and

$$\nu_{2i} = \frac{1}{\nu_{2i}} \{\nu_{1i}v_{1i}+\underline{\mu}_{1i}'\Sigma_{1i}^{-1}\underline{\mu}_{1i}+(n-3)s_i^2+\hat{\underline{\theta}}_i'W_i'W_i\hat{\underline{\theta}}_i$$

$$+ \underline{\mu}_{2i}'\Sigma_{2i}^{-1}\underline{\mu}_{2i}\}$$

Assuming $p_{1i} = 3$ for all i, we get $p_{2i} = 3$ and $\nu_{2i} = n + \nu_{1i}$.

Let the quantity in (3.7) equal A_i. Then

$$\int_{\Theta} A_i d\underline{\theta} = B_i ,$$

where

$$B_i = (2\pi)^{3/2}|\Sigma_{1i}|^{-1/2}|\Sigma_{2i}|^{1/2} h^{\nu_{1i}/2-1} \exp\{-h\nu_{2i}v_{2i}/2\}$$

$$\times (\nu_{1i}v_{1i}/2)^{\nu_{1i}/2}[\Gamma(\nu_{1i}/2)]^{-1}P(\gamma=c_i) .$$

Now, let

$$\int_H B_i \, dh = c_i ,$$

where

$$c_i = \frac{(2\pi)^{3/2} |\Sigma_{1i}|^{-1/2} |\Sigma_{2i}|^{1/2} (\nu_{1i} v_{1i}/2)^{\nu_{1i}/2} \Gamma(\nu_{1i}/2) P(\gamma = c_i)}{\Gamma(\nu_{1i}/2)(\nu_{2i} v_{2i}/2)^{\nu_{2i}/2}} .$$

Using the fact that

$$(T')^{-1} \sum_{i=1}^{k} c_i = 1,$$

we get

$$(3.9) \qquad T' = (2\pi)^{3/2} \sum_{i=1}^{k} \frac{|\Sigma_{1i}|^{-1/2} |\Sigma_{2i}|^{1/2} (\nu_{1i} v_{1i}/2)^{\nu_{1i}/2} \Gamma(\nu_{2i}/2) P(\gamma = c_i)}{\Gamma(\nu_{1i}/2)(\nu_{2i} v_{2i}/2)^{\nu_{2i}/2}}$$

In much of the future use of this term, the multiplier $(2\pi)^{3/2}$ will cancel out. For this reason we define $T = (2\pi)^{-3/2} T'$ and will use the constant T throughout.

Using the results obtained above, the joint posterior distribution consists of k parts, each part being of the form of (3.7) divided by (3.9). These distributions have parameters defined by (3.8).

3.4 Marginal and Conditional Posterior Distributions

In analysis of the posterior distributions, we are not really interested in the joint posterior distributions. The information which is most straightforward is contained in the marginal posterior probabilities concerned with the location of the join point, and the conditional distribution for $\underline{\theta}$ and h, given $\gamma = c_i$. For any possible value of γ for which we stated a prior probability, we can determine the conditional posterior density for $\underline{\theta}$ and h for that particular value of γ, and also determine the marginal posterior probability that the join point actually occurs at that point. $g_{2i}(\underline{\theta}, h/W(c_i), \underline{Y})$ indicates the posterior density function of $\underline{\theta}$ and h, given $\gamma = c_i$. Although the conditioning on $\gamma = c_i$ is not explicit, it is contained in the matrix $W(c_i)$. We use the notation $P(\gamma=c_i/X, \underline{Y})$ for the marginal posterior probability that $\gamma = c_i$. This posterior probability is not dependent on $W(c_i)$, so the dependency on the design matrix is shown by using X instead.

3.4.1 Marginal Posterior Probabilities of the Join Point γ

The marginal posterior probability that the unknown join point γ equals c_i, namely $P(\gamma=c_i/X, \underline{Y})$, is, by integrating the joint posterior distribution over the domain of $\underline{\theta}$ and h, equal to

$$(3.10)\qquad P(\gamma=c_i/X,Y)$$

$$= \int_H \int_\Theta g_{2i}(\underline{\theta},h/W(c_i),Y)P(\gamma=c_i/X,Y)\underline{d\theta}dh$$

$$= \frac{|\Sigma_{1i}|^{-1/2}|\Sigma_{2i}|^{1/2}(\nu_{1i}v_{1i}/2)^{\nu_{1i}/2}\underline{\Gamma}(\nu_{2i}/2)p_{1i}}{\underline{\Gamma}(\nu_{1i}/2)(\nu_{2i}v_{2i}/2)^{\nu_{2i}/2}/T}.$$

Assume the prior information on the parameter h is the same for the k different possible join points. The same prior information on h for the k different cases is equivalent to $\nu_{1i} = \nu_1$ and $v_{1i} = v_1$, for $i = 1,\ldots,k$. In this case (3.10) reduces to

$$(3.11)\qquad P(\gamma=c_i/X,\underline{Y}) = \frac{|\Sigma_{1i}|^{-1/2}|\Sigma_{2i}|^{1/2}p_{1i}/(\nu_2 v_{2i}/2)^{\nu_2/2}}{\sum_{i=1}^{k}\{|\Sigma_{1j}|^{-1/2}|\Sigma_{2j}|^{1/2}p_{1j}/(\nu_2 v_{2j}/2)^{\nu_2/2}\}}.$$

In most situations the posterior probabilities for the various possible values of the join point are given by either (3.10) or (3.11).

3.4.2 Conditional Distribution of $\underline{\theta}$ and h for a Fixed Value of γ

Since we know the form of $P(\gamma=c_i/X,\underline{Y})$ for any c_i, we can get $g_{2i}(\underline{\theta},h/W(c_i),\underline{Y})$ by factoring the marginal probability out of the joint posterior distribution at that c_i. The joint posterior distribution is given by (3.7) divided by (3.9). Factoring out the

marginal posterior probability that $\gamma = c_i$ we find

$$(3.12)\quad g_{2i}(\underline{\theta}, h/W(c_i), \underline{Y})$$

$$= (2\pi)^{-3/2} h^{3/2} |\Sigma_{2i}|^{-1/2} \exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu}_{2i})\Sigma_{2i}^{-1}(\underline{\theta}-\underline{\mu}_{2i})\}$$

$$\times h^{\nu_{2i}/2-1} \exp\{\nu_{2i}v_{2i}/2\}(\nu_{2i}v_{2i}/2)^{\nu_{2i}/2}[\Gamma(\nu_{2i}/2)]^{-1} .$$

This density function is the joint posterior density function for $\underline{\theta}$ and h, conditioned on the fact that $\gamma = c_i$. There are k of these posterior density functions, one for each possible value of γ. But $g_{2i}(\underline{\theta}, h/W(c_i), \underline{Y})$ itself can be factored into different conditional and marginal density functions. We know that the density function can be written in either of the two following forms.

$$(3.13)\quad g_{2i}(\underline{\theta}, h/W(c_i), \underline{Y}) = f_{2i}(\underline{\theta}/h, W(c_i), \underline{Y})b_{2i}(h/W(c_i), \underline{Y}) ,$$

$$(3.14)\quad g_{2i}(\underline{\theta}, h/W(c_i), \underline{Y}) = m_{2i}(\underline{\theta}/W(c_i), \underline{Y})q_{2i}(h/\underline{\theta}, W(c_i), \underline{Y}) .$$

First consider (3.13). In this expression the joint posterior density function, conditioned on a possible value of the join point, is factored into the conditional density function of $\underline{\theta}$ given h, and the marginal posterior density function of h. This is called a marginal distribution even though both the conditional and marginal distributions are also conditioned on the value c_i. By integrating $\underline{\theta}$ out of

(3.12), we get

$$(3.15)\qquad b_{2i}(h/W(c_i),\underline{Y}) = \frac{h^{\nu_{2i}/2-1}\exp\{-h\nu_{2i}v_{2i}/2\}(\nu_{2i}v_{2i}/2)^{\nu_{2i}/2}}{\Gamma(\nu_{2i}/2)}$$

which is similar in form to the marginal prior distribution assumed for h when γ is known. The density function of that distribution is given by (2.23). If we factor the above expression out of (3.12), the remainder is the conditional density function of $\underline{\theta}$, given h and $\gamma = c_i$. This is

$$(3.16)\qquad f_{2i}(\underline{\theta}/h, W(c_i), \underline{Y}) = (2\pi)^{-3/2}|\Sigma_{2i}|^{-1/2}\exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu}_{2i})'\Sigma_{2i}^{-1}(\underline{\theta}-\underline{\mu}_{2i})\}\ .$$

This density function is the density function of a trivariate normal distribution with mean vector $\underline{\mu}_{2i}$, and dispersion matrix $h^{-1}\Sigma_{2i}$. For the special case where the precision is assumed known, there obviously is no marginal distribution for h, but the posterior density of $\underline{\theta}$ for a given $\gamma = c_i$ is merely that given by (3.16) for the known value of h.

Using (3.14) the joint posterior density function is factored into a conditional distribution of h given $\underline{\theta}$, and a marginal posterior distribution for $\underline{\theta}$. As with the other functions, there are k sets of these posterior functions. By integrating the joint posterior density function over the domain of h, we get

$$(3.17)\qquad m_{2i}(\underline{\theta}/W(c_i),\underline{Y})$$

$$= \frac{(2\pi)^{-3/2}|\Sigma_{2i}|^{-1/2}(\nu_{2i}v_{2i}/2)^{\nu_{2i}/2}\,\Gamma(\frac{1}{2}(\nu_{2i}+3))}{\Gamma(\nu_{2i}/2)\{\frac{1}{2}[\underline{\theta}-\underline{\mu}_{2i})'\Sigma_{2i}^{-1}(\underline{\theta}-\underline{\mu}_{2i})+\nu_{2i}v_{2i}]\}^{(\nu_{2i}+3)/2}}$$

$$(3.18)\qquad = \frac{(\pi)^{-3/2}|\Sigma_{2i}^{-1}/v_{2i}|\,\nu_{2i}^{\nu_{2i}/2}\,\Gamma((\nu_{2i}+3)/2)}{\Gamma(\nu_{2i}/2)[(\underline{\theta}-\underline{\mu}_{2i})'\Sigma_{2i}^{-1}/\nu_{2i}(\underline{\theta}-\underline{\mu}_{2i})+\nu_{2i}]^{(\nu_{2i}+3)/2}} .$$

This density function is the density function of a multivariate Studnet's -t distribution as originally defined by Dunnett and Sobel (5). The mean vector of this distribution is $\underline{\mu}_{2i}$ and the dispersion matrix is equal to

$$(3.19)\qquad \Sigma_{2i}v_{2i}\left(\frac{\nu_{2i}}{\nu_{2i}-2}\right) .$$

By factoring (3.17) out of the joint posterior density function, we find

$$(3.20)\qquad q_{2i}(h/\underline{\theta}, W(c_i), Y) = \frac{h^{\nu_{3i}/2-1}(\nu_{3i}v_{3i}/2)^{\nu_{3i}/2}\exp\{-h\nu_{3i}v_{3i}/2\}}{\Gamma(\nu_{3i}/2)}$$

where

$$\nu_{3i} = \nu_{2i} + 3,$$

and

$$v_{3i} = \frac{1}{\nu_{3i}}\left[(\underline{\theta}-\underline{\mu}_{2i})'\Sigma_{2i}^{-1}(\underline{\theta}-\underline{\mu}_{2i})+\nu_{2i}v_{2i}\right] .$$

This density is also the same form as the prior given by (2.23). This result illustrates one of the reasons why the conjugate prior distributions are used in this analysis. The form of the density function has remained unchanged from prior belief to posterior belief, allowing easy interpretation of the change in belief which the likelihood function has caused. The change in the values of the parameters is all that need be considered.

3.5 Analysis of the Posterior Distribution

The main purpose in regression analysis is to increase the information about the unknown parameters of the underlying model. In the Bayesian approach being used here, all of the information about these parameters is contained in the posterior distributions. In the preceding section the posterior distributions are expressed in many alternate forms. This allows a person to choose that form which is most applicable to the situation under study. However, at this time we will look more closely at those parts of the posterior distribution that are of primary importance in most situations. The posterior distributions which we will examine more closely are the marginal posterior probabilities of the join point, and the marginal posterior distribution of $\underline{\theta}$. This last distribution is actually dependent on the value of γ, but we call it a marginal distribution since it will not depend on h.

3.5.1 Marginal Posterior Probabilities of the Join Point

Assume that $\nu_{1i} = \nu_1$ and $v_{1i} = v_1$, for $i = 1, \ldots, k$. Then, by (3.11), the posterior probability that $\gamma = c_i$ is proportional to

$$(3.20) \qquad |\Sigma_{1i}|^{-1/2}|\Sigma_{2i}|^{1/2}P(\gamma=c_i)(\nu_2 v_{2i}/2)^{-\nu_2/2}, \quad i = 1, \ldots, k.$$

In the methods based on least squares, the estimate of the join point is that estimate of γ belonging to the set of estimates which yields the smallest value of SSE. If normally distributed errors are assumed, maximum likelihood yields estimates that also minimize SSE. Using the Bayesian approach an experimenter would say the most likely value for the join point, or the "best posterior estimate," is that value of c_i which has the largest marginal posterior probability. How does this compare to the least squares approach? In (3.20) there are five parameters, two which are solely functions of the prior distribution, namely Σ_{1i} and $P(\gamma=c_i)$, and three which depend on both the prior distribution and the likelihood. One of these three, ν_2, depends on the likelihood function only through the sample size. By our assumption this parameter is also the same for all of the possible values of the join point. The other two parameters rely heavily on values calculated from the sample, namely $\hat{\underline{\theta}}_i$ and s_i^2. These sample estimates of the parameters change for the various join points.

The parameter v_{2i} corresponds to a posterior estimate of σ^2, assuming c_i is the actual join point. It is not an unbiased estimate. It is the term, though, which most closely corresponds to SSE term used in least squares as a criterion. Considering (3.20), we see that, all other four parameters being equal, if one possible join point yields a value for v_2 that is smaller than the value of v_2 for a second possible join point, then the posterior probability for the first join point will be larger. In this sense there is some relationship between the least squares criterion and the marginal posterior probabilities.

Very seldom for any two possible join points will all parameters with the exception of v_2 be the same. In most cases $|\Sigma_{1i}|$ and $|\Sigma_{2i}|$ come into play. These determinants are the generalized variance of the prior and posterior conditional distributions on $\underline{\theta}$, respectively. Box and Draper (4) discuss the properties of the generalized variance. We will not go into the properties here, since in all examples considered, the product of the determinants used in (3.20) is relatively stable for reasonable changes in the join point. In the example considered below, this property will be illustrated.

$P(\gamma=c_i)$, the prior belief that $\gamma = c_i$, also plays a role in the determination of the posterior marginal probabilities. Depending on the prior knowledge of the situation, this parameter may be very important. It should be brought out that as the sample size increases, the term v_{2i} becomes increasingly dominant, and thus a criterion

based on the largest posterior probability will approach the least squares criterion in the limit. However, for intermediate sample sizes the posterior probabilities take much more into account.

The above discussion is not meant to indicate that the possible join point with the largest posterior marginal probability should be chosen, and all of the other possible join points ignored. The information on the join points should be used with the posterior distributions on $\underline{\theta}$ to give the experimenter a good idea about the most likely models for the problem under study.

3.5.2 The Posterior Distribution of $\underline{\theta}$, Given $\gamma = c_i$

As previously determined, the marginal posterior distribution of $\underline{\theta}$ for $\gamma = c_i$, is a multivariate Student's-t distribution. The density is given by (3.18). As in the case of a multivariate normal distribution, all of the information is contained in the mean vector and the dispersion matrix. If we again assume $\nu_{1i} = \nu_1$, the mean vector is $\underline{\mu}_{2i}$, and the dispersion matrix is equal to (3.19). In any situation the vector $\underline{\mu}_{2i}$ will be of principal interest. This vector contains the posterior means of the three parameters θ_1, θ_2, θ_3, or, in terms of the original model, α_0, α_1, $\beta_1 - \alpha_1$. This vector should be considered for all situations where the posterior probability of the join point is large. It can indicate many things concerning the model for the data which we are studying. One example is the case

where the posterior mean of θ_3 is approximately zero. This gives a good indication that the underlying model might actually be just one straight line segment.

As is the case for multivariate normal distributions, the marginal distribution of any subset of variables having a multivariate-t distribution also has a multivariate (or univariate) Student's-t distribution. Using this fact we can look at the marginal posterior distribution of any of the three parameters. Recently, several papers have appeared in which tables of the percentage points of the trivariate-t distribution have appeared. One especially good table is that given by Trout and Chow (18). Using these tables regions can be found where the posterior probability content for the three unknown values in $\underline{\theta}$ is, say 95 percent. There also exists tables for the bivariate case.

3.6 Example

The data for this example is the same data as that studied in Section 1.3. The observations are given at the first part of that section, and a plot of the data is given in Figure 1. In Chapter 1 we showed that the two-phase linear regression line yielded a good fit when compared to other possible models. A least squares approach was used to fit the model. In this example, we will analyze the same data using the Bayesian approach which we have developed.

As in example 2.4, we will start the analysis with the likelihood function, then consider the prior distribution, and lastly, consider the posterior distribution. However, before considering any of the three parts of Bayesian analysis, some attention must be given to "discretizing" the parameter γ. For this example five potential join points are considered. There are increments of equal size between the possible values, two which occur at x-observations, and three which occur in three distinct intervals. The values considered are: 5.5, 6.0, 6.5, 7.0, 7.5. The theory imposes no restriction on the spacing of possible values of γ, and a much larger or smaller number of possible join points can be considered in a given problem. It is felt that these five points will adequately illustrate all pertinent concepts.

(1) The likelihood function. Since there are five potential join points, there are five different likelihood functions. Denote $W_i = XA(c_i)$, $i = 1, \ldots, 5$. We let $i = 1$ indicate the possible join point 5.5, $i = 2$ indicate the possible join point 6.0, and so forth. Let us look at W_1.

$$W_1 = XA(5.5) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 1 & 5 & 0 & 0 \\ 0 & 0 & 1 & 6 \\ 0 & 0 & 1 & 7 \\ 0 & 0 & 1 & 8 \\ 0 & 0 & 1 & 9 \\ 0 & 0 & 1 & 10 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & -5.5 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 4 & 0 \\ 1 & 5 & 0 \\ 1 & 6 & .5 \\ 1 & 7 & 1.5 \\ 1 & 8 & 2.5 \\ 1 & 9 & 3.5 \\ 1 & 10 & 4.5 \end{bmatrix}$$

The forms for W_i, $i = 2, \ldots, 5$ are similar, but the elements in the last column change from one potential join point to another. From W_i we get the more important matrix $W_i'W_i$ and its inverse. As an example consider these matrices for $i = 1$.

$$W_1'W_1 = \begin{bmatrix} 10 & 55 & 12.5 \\ & 385 & 110 \\ & & 41.5 \end{bmatrix}, \text{ and } (W_1'W_1)^{-1} = \begin{bmatrix} .917 & -.217 & .300 \\ & .062 & -.100 \\ & & .200 \end{bmatrix}.$$

For $W_i'W_i$ the upper corner 2×2 main submatrix is the same for all values of the join point. However, the bottom row and the last column depend on the value of the join point. We will not show the other four matrices and their inverses because of the similarity to those given below.

The other information obtained from the likelihood function which we use in the posterior distribution is contained in $\hat{\underline{\theta}}_i$, and the SSE. These values are given in Figure 6. It should be pointed out here that the SSE for $\gamma = 6.5$ is within .02 of SSE for the optimal least squares estimate found in example 1.3. The above information,

together with the matrices $W_i'W_i$ and their inverses contain all the information in the data.

Estimate	Potential Join Point $\gamma = 5.5$	$\gamma = 6.0$	$\gamma = 6.5$	$\gamma = 7.0$	$\gamma = 7.5$
$\hat{\underline{\theta}}_i$	0.26	0.59	0.82	1.28	1.58
	2.67	2.51	2.41	2.24	2.14
	-2.34	-2.39	-2.61	-2.68	-3.01
SSE	7.64	5.65	5.05	6.41	8.51

Figure 6. Estimates based on likelihood function.

(2) The prior distribution. As mentioned in Section 1.3, the data in this example was generated from a specific model. We will use this information in our prior, varying the degree of belief which we have in this knowledge. Therefore,

$$\underline{\mu_1(\phi)} = \begin{pmatrix} 2.5 \\ 2 \\ 12.25 \\ .5 \end{pmatrix}, \quad \text{and} \quad v_1 = 1. \tag{3.21}$$

Using these values we will derive the various constrained prior parameters.

First consider Σ_1. As in example 2.4 we start by assuming an unconstrained prior dispersion matrix, $\Sigma_1(\phi) = (10)^k I_4$, $k = -1, 0, 1, 2, 3$. Therefore there are five possible constrained prior dispersion matrices for each possible value of the join point. By

varying k, we vary our belief in the prior vector $\underline{\mu_1(\phi)}$. By (3.5)

$$\Sigma_{1i} = B(I-\Sigma_1(\phi)\underline{\delta(c_i)}\{\underline{\delta'(c_i)}\Sigma_1(\phi)\underline{\delta(c_i)}\}^{-1}\underline{\delta'(c_i)}\Sigma_1(\phi)B',$$

where $\Sigma_1(\phi)$ is the unconstrained prior dispersion matrix and Σ_{1i} is the constrained prior dispersion when $\gamma = c_i$. Table 3.1 gives the values of Σ_{1i}, $i = 1,\ldots,5$ for the case when $\Sigma_1(\phi) = I_4$. All of the other prior dispersion matrices can be obtained by multiplying these matrices by an appropriate power of ten.

Table 3.1. Constrained prior dispersion matrices for various possible join points when $\Sigma_1(\phi) = I_4$.

	.984	-.088	.176
$\gamma = 5.5$		.516	-.032
			.064
	.986	-.081	.162
$\gamma = 6.0$		.513	-.027
			.054
	.988	-.075	.150
$\gamma = 6.5$		.512	-.023
			.046
	.990	-.070	.140
$\gamma = 7.0$		.510	-.020
			.040
	.991	-.065	.131
$\gamma = 7.5$		.509	-.017
			.035

For each of the five possible join points, we also have to determine the constrained prior mean vector for $\underline{\theta}$, namely $\underline{\mu}_{1i}$. Since we assumed $\Sigma_1(\phi) = (10)^k I_4$, for the formula for the constrained prior mean vector is given by (2.36). We use the unconstrained prior mean vector given by (3.21). The values obtained for the constrained prior mean vectors are given in Figure 7.

$\gamma = 5.5$	$\gamma = 6.0$	$\gamma = 6.5$	$\gamma = 7.0$	$\gamma = 7.5$
$\begin{pmatrix} 2.52 \\ 2.13 \\ -1.76 \end{pmatrix}$	$\begin{pmatrix} 2.51 \\ 2.06 \\ -1.62 \end{pmatrix}$	$\begin{pmatrix} 2.50 \\ 2.00 \\ -1.50 \end{pmatrix}$	$\begin{pmatrix} 2.49 \\ 1.95 \\ -1.39 \end{pmatrix}$	$\begin{pmatrix} 2.49 \\ 1.90 \\ -1.30 \end{pmatrix}$

Figure 7. Prior constrained mean vector for $\underline{\theta}$ at the five possible join points.

For all possible join points, we let $v_1 = 1$ and let ν_1 range over the values 1, 10, and 50. For the prior beliefs in the various join points, namely $P(\gamma=c_i)$, we assume they are all equal and $P(\gamma=c_i) = .2$, $i = 1, \ldots, 5$.

(3) The posterior distribution. In analyzing the posterior distribution we will concentrate on the posterior parameter vectors $\underline{\mu}_{2i}$, and the posterior marginal probabilities of the various possible join points. The other parameters will now be briefly mentioned.

The posterior parameter $\nu_{2i} = 10 + \nu_1$, where $\nu_1 = 1, 10, 50$. Therefore, ν_{2i} is one of three values, 11, 20, or 60. The

posterior dispersion matrices Σ_{2i} were calculated for all possible combinations of join points and prior dispersion matrices. However, for any given prior dispersion matrix, these posterior dispersion matrices proved to be fairly similar and will not be discussed further. In given situations these matrices are necessary if the analysis is to include study of the probability content of a region, or if the posterior distributions of a subset of the vector $\underline{\theta}$ is to be studied. The posterior parameter v_{2i} also will not be explicitly studied. Its major importance lies in its contribution to the marginal posterior probabilities of the join point.

Let us now consider the posterior marginal probabilities of the join point. These probabilities are given in Table 3.2. The formula for the posterior probability is given by (3.11). In this example the prior probability was assumed to be the same for each of the five possible join points. The product $|\Sigma_{1i}|^{-1/2}|\Sigma_{2i}|^{1/2}$ proved to be of negligible effect in this example in determining the posterior probability. Its value was relatively stable for changing values of the join point when the prior dispersion matrix was held constant. Figure 8 gives the values of this product for the case when $\Sigma_1(\phi) = (10)^{-1}I_4$. Therefore, the major influence in the posterior marginal probabilities of the join point comes from the term v_{2i}. However it does not correspond directly to SSE, which, by Figure 6, would pick $\gamma = 6.5$.

Table 3.2. Posterior probabilities for location of join point, γ.

Join Point	Value of Prior Dispersion Matrix, $\Sigma_1(\phi)$				
	$(10)^{-1}I$	I	$10I$	$(10)^2 I$	$(10)^3 I$
a) $\nu_1 = 1$ $(\nu_2 = 11)$					
$\gamma = 5.5$	.116	.171	.150	.081	.068
$\gamma = 6.0$	.424	.490	.470	.323	.284
$\gamma = 6.5$	.324	.254	.304	.446	.466
$\gamma = 7.0$	.114	.071	.067	.129	.151
$\gamma = 7.5$	.022	.014	.009	.021	.031
b) $\nu_1 = 10$ $(\nu_2 = 20)$					
$\gamma = 5.5$	.114	.175	.165	.100	.087
$\gamma = 6.0$	.435	.485	.441	.315	.285
$\gamma = 6.5$	.327	.260	.308	.407	.416
$\gamma = 7.0$	.107	.070	.076	.150	.171
$\gamma = 7.5$	.017	.010	.010	.028	.041
c) $\nu_1 = 50$ $(\nu_2 = 60)$					
$\gamma = 5.5$	.110	.178	.172	.113	.098
$\gamma = 6.0$	.448	.484	.427	.313	.280
$\gamma = 6.5$	.332	.263	.311	.380	.388
$\gamma = 7.0$	.098	.068	.082	.160	.184
$\gamma = 7.5$	.012	.007	.008	.034	.050

$\gamma = 5.5$	$\gamma = 6.0$	$\gamma = 6.5$	$\gamma = 7.0$	$\gamma = 7.5$
.2039	.2035	.2031	.2026	.2021

Figure 8. Calculated values for $|\Sigma_{1i}|^{-1/2}|\Sigma_{2i}|^{1/2}$ when $\Sigma_1(\phi) = (10)^{-1}I_4$.

In analyzing the posterior probabilities for the various possible join points, $\gamma = 6.0$ and $\gamma = 6.5$ are the points which deserve the most consideration. In nearly all cases these two points account for more than 70% of the posterior probability. For the situations where the belief in the prior mean is moderate to very strong, $\gamma = 6.0$ has the highest posterior probability. In the two cases where the prior is weak, the posterior probability is greatest for $\gamma = 6.5$. Both of these results hold for the three values of ν_1. The question might be raised why $\gamma = 6.5$ is not preferred in all cases, since that is the value for the model which generated the data. However, the least squares estimate for the slope of the second line when constrained to join at 6.5 is approximately (-.5), while the slope is (.5) for the actual model. This large difference comes about because of the huge deviations of some of the generated observations near the join point. However, these results do show that at $\gamma = 6.0$ the prior constrained mean vector most closely agrees with the constrained least squares estimates, when compared with the results at the other join points.

The other set of posterior parameters are $\underline{\mu}_{2i}$. The values of these posterior mean vectors are given in Table 3.3. We will not discuss the different ways to analyze these means, because much depends on the purpose of the analysis. However, some things should be noted. As in example 2.4, as the determinant of the dispersion matrix increases, the posterior mean vector ranges from values close

to the prior constrained mean vector to values close to the constrained least squares estimates. This occurs as k increases. In this example, also, the posterior mean vectors at $\gamma = 6.0$ and $\gamma = 6.5$ are the only ones that need be considered. The posterior probabilities indicate that the other three cases are not very likely.

Table 3.3. Posterior mean vectors of $\underline{\theta}$ for the five possible join points and five different prior dispersion matrices.

Join Point	Prior Dispersion Matrix, $\Sigma_1(\phi)$				
	$(10)^{-1}I$	I	$10I$	$(10)^2 I$	$(10)^3 I$
$\gamma = 5.5$	2.302	1.376	0.503	0.294	0.269
	2.218	2.402	2.602	2.659	2.666
	-1.801	-1.974	-2.234	-2.330	-2.342
$\gamma = 6.0$	2.327	1.612	0.865	0.629	0.597
	2.089	2.232	2.422	2.498	2.508
	-1.655	-1.827	-2.185	-2.362	-2.388
$\gamma = 6.5$	2.368	1.876	1.219	0.883	0.828
	1.988	2.088	2.275	2.391	2.411
	-1.529	-1.689	-2.183	-2.543	-2.604
$\gamma = 7.0$	2.417	2.168	1.711	1.355	1.287
	1.898	1.949	2.090	2.216	2.240
	-1.410	-1.534	-2.059	-2.570	-2.670
$\gamma = 7.5$	2.478	2.450	2.162	1.711	1.593
	1.830	1.840	1.941	2.096	2.136
	-1.310	-1.395	-1.942	-2.764	-2.978

4. BAYESIAN ANALYSIS OF TWO-PHASE REGRESSION WITH VAGUE PRIOR KNOWLEDGE

One of the primary criticisms of Bayesian inference in general, and the method developed in the two preceding chapters in particular, is the necessity of formulating the prior distribution for the problem. In the analysis which was developed in the preceding chapters, the general form of the prior distribution is known, but up to 5 parameters for this distribution must be specified. Although the examples in the previous chapters show that weak prior knowledge can be considered by varying Σ_1 and ν_1, there is another alternative to this method. This alternative method is the use of "vague priors." These priors are not distributions in the proper sense, but are functions that have come to represent situations in which the experimenter desires the posterior distribution on a parameter to be strictly a function of the information contained in the sample. In this chapter, we will first show the forms of the "vague priors" which we will assume for the various parameters, and then develop the posterior distributions for the various cases.

4.1 Vague Priors on the Parameters

In this section we consider vague priors for h (or σ^2), for the vector of parameters $\underline{\theta}$, and for the location of the join point, γ.

4.1.1 Vague Prior on h

In the prior distribution for h, the two parameters which must be stipulated are v_1 and ν_1. The parameter v_1 is a prior estimate of σ^2, while the parameter ν_1 indicates the degree of belief which we have in the value of v_1. For this reason, a logical convention for situations where there is vague prior knowledge is to set $\nu_1 = 0$. When using this restriction the marginal distribution of h is proportional to h^{-1}, and

$$(4.1) \quad g_{1i}(\underline{\theta}, h/\gamma=c_i) = h^{p_{1i}/2} \exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu}_{1i})'\Sigma_{1i}^{-1}(\underline{\theta}-\underline{\mu}_{1i})\}h^{-1} .$$

In most situations if a vague prior is assumed for an unknown parameter at one join point, a vague prior will be assumed for all join points.

4.1.2 Vague Prior on $\underline{\theta}$

In the case where there is vague prior knowledge on $\underline{\theta}$, an obvious result is that the variance of the prior mean vector is very large. As was evident in the examples of the previous two chapters, by letting $\Sigma_1(\phi)$, or equivalently Σ_1, get very large, we reduce the effect of the prior distribution of $\underline{\theta}$ on the posterior distribution of $\underline{\theta}$. Therefore, for a vague prior on $\underline{\theta}$, we assume Σ_1 is

infinitely large, or equivalently, $\Sigma_1^{-1} = 0$, and $p_{1i} = 0$. Resulting from this, the prior density function $g_{1i}(\underline{\theta}, h/\gamma=c_i)$ is proportional to

$$h^{-\nu_{1i}/2-1} \exp\{-h\nu_{1i}v_{1i}/2\} . \tag{4.2}$$

4.1.3 Vague Prior on $\underline{\theta}$ and h

In this case ν_{1i}, Σ_{1i}^{-1} and p_{1i} are all set equal to zero, and we use the convention that

$$g_{1i}(\underline{\theta}, h/\gamma=c_i) \propto h^{-1} . \tag{4.3}$$

4.1.4 Vague Prior on γ

If we assume that the join point can occur at k different values, we assign the probability k^{-1} for each of the possible values.

In many cases vague prior knowledge on γ means not only that we are unsure about the prior probabilities of the various possible join points, but also that we are unsure about the values which γ can assume. In a situation such as this, one can proceed in the following way. Find the best least squares estimate of the join point for each interval by one of the classical methods such as Hudson's. For each of these estimates, the experimenter assumes the same prior

probability of the join point, and proceeds with the Bayesian analysis as if these values were his prior idea about the location of the join point. If there are also vague priors on h and $\underline{\theta}$, we are assuming no prior information and letting the sample wholly determine the posterior distributions.

4.2 Posterior Distributions When Using Vague Priors

Throughout this section and the previous section, h is assumed unknown. The situation where h is known is merely a particular case of the problem considered here, and the results for that case are obvious. In this section the posterior distributions are given for the various vague priors previously stated. For all of those priors, the same form of the posterior distribution results. That is, in all cases the joint posterior distribution is equal to (3.6), or

$$(4.4)\quad P(\gamma=c_i/X,Y)(2\pi)^{-3/2}|\Sigma_{2i}|^{-1/2}h^{3/2}\exp\{-\frac{h}{2}(\underline{\theta}-\underline{\mu}_{2i})'\Sigma_{2i}^{-1}(\underline{\theta}-\underline{\mu}_{2i})\}$$
$$\times h^{\nu_{2i}/2-1}\exp\{-h\nu_{2i}v_{2i}/2\}(\nu_{2i}v_{2i}/2)^{\nu_{2i}/2}[\Gamma(\nu_{2i}/2)]^{-1}$$

where $P(\gamma=c_i/X,\underline{Y})$ is given by (3.10). However the values of the parameters of this posterior distribution for the various cases of vague priors are different, and primary consideration will be given to the expression of these parameters in the different cases. In all

situations of vague prior knowledge, the expression of the likelihood used in Chapter 3, namely (3.3), remains the same.

The posterior distributions for the case where a vague prior is placed on γ will not be considered separately. In these cases, $P(\gamma=c_i) = k$, for all i, and this term simply falls out of $P(\gamma=c_i/X,\underline{Y})$.

4.2.1 Posterior Distribution for a Vague Prior on h

By combining the expression in (4.1) with the likelihood function given by (3.3), we get a joint posterior distribution whose density function is given by (4.4), where the parameters are defined as follows:

$$(4.5) \qquad \Sigma_{2i}^{-1} = \Sigma_{1i}^{-1} + W_i'W_i, \qquad \underline{\mu}_{2i} = \Sigma_{2i}(\Sigma_{1i}^{-1}\underline{\mu}_{1i} + W_i'W_i\hat{\underline{\theta}}_i)$$

$$p_{3i} = 3, \qquad \nu_{2i} = n$$

and

$$v_{2i} = \frac{1}{\nu_{2i}} \{(n-3)s_i^2 + \hat{\underline{\theta}}_i'W_i'W_i\hat{\underline{\theta}}_i + \underline{\mu}_{1i}'\Sigma_{1i}^{-1}\underline{\mu}_{1i} - \underline{\mu}_{2i}'\Sigma_{2i}\underline{\mu}_{2i}\}$$

The vague prior on h has no effect on the posterior parameters Σ_{2i}^{-1}, $\underline{\mu}_{2i}$, and p_{2i}. These parameters are wholly concerned with the posterior distribution of $\underline{\theta}$. However, the parameters v_{2i} and ν_{2i} have changed from the form given by (3.8), the parameters resulting when a proper prior is used. The posterior probability of the join point is still given by (3.11) with the parameters defined in

(4.5) being used in the formula.

4.2.2 Posterior Distribution When Assuming a Vague Prior on $\underline{\theta}$

Combine the expression given by (4.2) with the likelihood function and we get the same posterior distribution form as that given by (4.4), with the posterior parameters defined as follows:

$$\Sigma_{2i}^{-1} = W_i'W_i, \qquad \underline{\mu}_{2i} = \hat{\underline{\theta}}_i, \tag{4.6}$$

$$p_{2i} = 3, \qquad \nu_{2i} = n + \nu_{1i} - 3,$$

and

$$v_{2i} = \frac{1}{\nu_{2i}}\left[(n-3)s_i^2 + \hat{\underline{\theta}}_i W_i'W_i \hat{\underline{\theta}}_i + \nu_{1i}v_{1i} - \underline{\mu}_{2i}'\Sigma_{2i}^{-1}\underline{\mu}_{2i}\right]$$

$$= \frac{1}{\nu_{2i}}\left[(n-3)s_i^2 + \nu_{1i}v_{1i}\right].$$

In this situation all of the posterior parameters with the exception of p_{2i} have been significantly effected by the vague prior. The two parameters primarily concerned with $\underline{\theta}$, namely Σ_{2i}^{-1} and $\underline{\mu}_{2i}$, are functions of the sample alone. The posterior estimate of σ^2 is now merely a weighted linear combination of the prior estimate and the least squares estimate, with the weights being proportional to the prior degree of belief, and the degrees of freedom for error based on the sample. $P(\gamma=c_i/X,\underline{Y})$ is given by (3.10), but the parameters used in that formula are now those defined by (4.6).

4.2.3 Posterior Distribution When Assuming a Vague Prior on $\underline{\theta}$ and h

By combining h^{-1} and $P(\gamma=c_i)$ with the likelihood function, we again get a posterior distribution that has the form given by (4.4), where the posterior parameters are defined in the following way:

$$\Sigma_{2i}^{-1} = W_i'W_i, \qquad \underline{\mu}_{2i} = \hat{\underline{\theta}}_i, \tag{4.7}$$

$$p_{2i} = r(W_i'W_i) = 3, \qquad \nu_{2i} = n - 3,$$

and

$$v_{2i} = \frac{1}{\nu_{2i}}[(n-3)s_i^2] = s_i^2 .$$

All of the posterior parameters are functions only of the sample, and $\underline{\mu}_{2i}$ and v_{2i} correspond to the least squares estimates of $\underline{\theta}$ and σ^2, respectively.

In this case where there is no prior knowledge on any of the parameters except the join point, the posterior estimates of the join point is important to study. By inserting the parameters defined in (4.7) into (3.11), we get

$$P(\gamma=c_i/X,Y) = \frac{|W_i'W_i|^{-1/2}((n-3)s_i^2)^{-(n-3)/2}P(\gamma=c_i)}{\sum_{j=1}^{k}\{|W_j'W_j|^{-1/2}((n-3)s_j^2)^{-(n-3)/2}P(\gamma=c_i)\}} . \tag{4.8}$$

If it is assumed that the prior probability on all possible join points is the same, the posterior probability that the join points equals c_i is proportional to

$$|W_i'W_i|^{-1/2}((n-3)s_i^2)^{-(n-3)/2} .$$

As a criterion to choose the most likely join point, the above form is very close to Hudson's approach which solely uses the minimizing of s_i^2 as a criterion. For our Bayesian posterior with vague priors on all unknown parameters, a function of s_i^2 is multiplied by the generalized variance raised to the $(-1/2)$ power. However, as in example 3.6 the generalized variance is usually relatively stable over any interval which we would consider. Therefore, the Bayesian approach is very similar to the least squares approach for the situation where vague priors are assumed.

If this vague prior approach is considered for example 3.6, the posterior probabilities of the five possible join points are

$\gamma = 5.5$	$\gamma = 6.0$	$\gamma = 6.5$	$\gamma = 7.0$	$\gamma = 7.5$
.088	.250	.400	.182	.080

If these points are ordered according to highest probability, the order is the same as if they are ordered according to smallest SSE.

5. SUMMARY AND CONCLUSIONS

In this investigation the two-phase linear regression model has been considered using a Bayesian approach. The method presented differs from previous methods in its ability to insert prior knowledge concerning the parameters of the underlying model into the analysis.

In Chapter 2 the Bayesian approach was developed for the situation where the join point is known. It was shown that if a continuity condition is imposed on the model, the likelihood function should be redefined in order to explicitly contain the condition within the function. The model starts out with four parameters, excluding the variance and join point. In order to impose the condition, only three parameters can be explicitly used. There are many ways to do this. The transformation used in this investigation resulted in the intercept and slope parameters of the first line segment, and the difference between the slope of the second line segment and the slope of the first line segment as the three parameters considered.

The natural conjugate prior distribution was assumed. Using a conjugate prior places the principal emphasis on the determination of the posterior parameters, since the form of the posterior distribution will be the same as the form of the prior distribution. A method was developed to find the constrained prior mean vector and dispersion matrix for any pair of unconstrained mean vector and dispersion

matrix. When the variance term is unknown, there are essentially four sets of prior parameters to determine. Two of these sets, $\underline{\mu}_{1i}$ and v_{1i}, are prior estimates of the unknown parameters of the model, while the other two sets, namely Σ_{1i} and ν_{1i}, indicate the degree of belief in the other prior estimates. The posterior distribution was derived, and the correspondence of the posterior parameters to functions of the sample and prior parameter was developed.

Chapter 3 considered the case where the join point is unknown. The methodology developed in Chapter 2 was extended to the situation where the join point assumes one of a finite number of points. The restriction of finiteness is made in order to summarize the data in sufficient statistics, and thus be able to continue to use a natural conjugate prior. Posterior distribution was developed, and the expression of the posterior marginal probability of the possible join points is shown. This posterior seems to be more complex than the ordinary criterion used in a two-phase linear regression, namely SSE.

In Chapter 4, the Bayesian approach for the case of vague prior knowledge was developed. The results correspond closely to the results obtained by a least squares approach.

In trying to analyze the results of the Bayesian approach, both positive and negative aspects can be considered. Let us first consider the negative aspects. In a situation where there is not prior information on any of the unknown parameters, very little information is

gained which is not available by the least squares approach. Also, if some prior information does exist, much more time must be spent in the analysis than in the least squares method. A more serious negative aspect from a theoretical point of view might be the "discretizing" of the possible values of the join point. However, in the example considered, this "discretizing" resulted in at least one value which was close, in a least squares sense, to the optimal classical estimate.

The positive aspects of the Bayesian approach are well worth considering. First, this approach gives a systematic method to consider different possible join points. The "discretizing" factor can be used to good advantage if more than one possible join point in a given interval is to be considered. Even with vague priors, the posterior information is in a form which is easy to interpret. Second, if any prior knowledge on any of the parameters exists, the Bayesian approach makes use of this knowledge, and through the parameter v_{2i} measures the agreement between the prior belief and the results of the sample. The use of a two-phase linear regression model on a set of data usually implies at least some prior conception or knowledge about the unknown parameters. None of the other methods in the literature are able to handle this knowledge within their framework. Even the approach of Bacon and Watts (1), though Bayesian in nature, can not handle prior knowledge. Their analysis assumes a vague prior, and in its present form cannot handle a proper prior

distribution. The third positive aspect is the degree of versatility that the prior parameters, especially the prior dispersion matrix, permits the experimenter to have. As was seen by the examples in Section 2.4 and Section 3.6, varying this parameter can result in a posterior mean vector nearly identical to the prior mean vector, or a posterior mean vector nearly identical to the vector of least squares estimates.

Although there are many other positive aspects, let us consider only one additional feature. This feature is the amount of information obtained. Although some people might prefer the results to be in a concise form, the two-phase linear regression model with unknown join point is a complex situation and not easy to analyze. The Bayesian approach yields a great deal of information in a form that is easy to analyze even from a frequentist standpoint. Through the use of the prior distributions many different situations can be considered, and various prior constraints can be imposed on the model.

In summary, it is felt that this approach is a reasonable alternative to the approaches already being used. It allows questions to be considered through the use of the prior distribution which could not be considered previously, and it gives a viable approach to a complicated problem.

BIBLIOGRAPHY

1. Bacon, D. and D. Watts. Estimating the transition between two intersecting straight lines. Biometrika 58:525-534. 1971.

2. Bellman, R. and R. Roth. Curve fitting by segmented straight lines. Journal of the American Statistical Association 64:1079-1084. 1969.

3. Bellman, R., B.G. Kashef, and R. Vasudevan. Splines via dynamic programming. Journal of Mathematical Analysis and Applications 38:471-479. 1972.

4. Box, M.J. and N.R. Draper. Factorial designs, the $|X'X|$ criterion, and some related matters. Technometrics 13:731-742. 1971.

5. Dunnett, C.W. and M. Sobel. A bivariate generalization of Student's t-distribution, with tables for certain special cases. Biometrika 41:153-169. 1954.

6. Fuller, W.A. Grafted polynomials as approximating functions. The Australian Journal of Agricultural Economics 13:35-46. 1969.

7. Gallant, A.R. and W.A. Fuller. Fitting segmented polynomial regression models whose join points have to be estimated. Journal of the American Statistical Association 68:144-147. 1973.

8. Hinkley, D.V. Inference about the intersection in two-phase regression. Biometrika 56:495-504. 1969.

9. Hinkley, D.V. Inference in two-phase regression. Journal of the Americal Statistical Association 66:736-743. 1971.

10. Hudson, D.J. Fitting segmented curves whose join points have to be estimated. Journal of the American Statistical Association 61:1097-1129. 1966.

11. McGee, V.E. and W.T. Carleton. Piecewise regression. Journal of the American Statistical Association 65:1109-1124. 1970.

12. McLaren, A.D. The fit of two straight lines when their intersection has a specified abscissa. Unpublished memorandum. University of Cambridge.

13. Quandt, R.E. The estimation of the parameters of a linear regression system obeying two separate regimes. Journal of the American Statistical Association 53:873-880. 1958.

14. Quandt, R.E. Tests of the hypothesis that a linear regression system obeys two separate regimes. Journal of the American Statistical Association 55:324-330.

15. Raiffa, H. and R. Schlaifer. Applied statistical decision theory. Boston, Harvard Business School, 1961. 365 p.

16. Robison, D.E. Estimates for the points of intersection of two polynomial regressions. Journal of the American Statistical Association 59:214-224. 1964.

17. Sprent, P. Some hypotheses concerning two phase regression lines. Biometrics 17:634-645. 1961.

18. Trout, J.R. and B. Chow. Table of the percentage points of the trivariate t-distribution with an application to uniform confidence bands. Technometrics 14:855-879. 1972.