# EMPIRICAL BAYES ESTIMATION OF RATES IN LONGITUDINAL STUDIES*

by

Siu L. Hui
Indiana University School of Medicine

James O. Berger
Purdue University

Technical Report #82-32

Department of Statistics
Purdue University

October 1982

*Siu L. Hui is Assistant Professor, Department of Medicine, Indiana University School of Medicine, Regenstrief Institute, 5th Floor, 1001 West Tenth Street, Indianapolis, IN 46202, and James O. Berger is Professor of Statistics, Department of Statistics, Purdue University, West Lafayette, IN 47907. Research was partially supported by PHS Grant No. S07-RR05443, PHS Grant No. R23-AG03423, and NSF Grant No. 8101670A1.

## Abstract

The usually irregular follow-up intervals in epidemiologic studies preclude the use of classical growth curve analysis. For short follow-up intervals, it is suggested that individual rates of change are useful for exploratory analysis. Empirical Bayes estimates of these rates of change are recommended for their smaller overall expected error than that of independent least squares estimates. An example of bone loss with age in women is given.

Keywords

Repeated measurements; growth curves; irregular follow-up; empirical Bayes estimates.

# 1. INTRODUCTION

Longitudinal studies are often undertaken to measure changes in the human body which may eventually lead to disease. Such studies require great effort and expense in order to make repeated measurements on the same subjects over many years; therefore, the valuable data collected should be fully utilized in subsequent statistical analyses.

Much of the earlier work on the analysis of repeated measurements has been developed in the study of growth curves, but it can be generalized to any dose-response relationship. Grizzle and Allen (1969) gave a comprehensive summary of those techniques developed from the generalized multivariate analysis of variance model of Pothoff and Roy (1964), which requires measurements on all subjects at each of several fixed time points. This constraint is a severe limitation for the analysis of epidemiologic studies in which subjects 1) often enter the study at different ages, 2) inevitably miss some scheduled visits, and/or 3) drop out of the study altogether. For longitudinal data with variable follow-up times, the choice of statistical methods is very limited. If one is mainly interested in testing for overall differences between groups of subjects, the nonparametric test by Zerbe and Walker (1977) is applicable.

In this paper, on the other hand, the main emphasis will be on model fitting and (local) estimation of individual response curves. The approach taken is essentially Bayesian or empirical Bayesian in nature, modeled after such papers as Lindley and Smith (1972), Efron and Morris (1973, 1975), Carter and Rolph (1974), Fearn (1975), Fay and Herriot (1979), Rolph, Williams, and Lee (1980), Rubin (1980), and Morris (1982).

An important feature of many epidemiologic studies is that the follow-up intervals of most of the individuals studied are sufficiently short that the response curves in the intervals can be approximated by straight lines, with slopes $\beta_i$. This will be seen to allow estimation of the individual slopes and of the mean population slope $\mu(t)$, without the need to specify functional forms for the true individual slopes over the entire age span of interest (a difficult and uncertain task). The empirical Bayes approach also gives individual slope estimates which are probably considerably more accurate than the least squares estimates, which may have large variances due to the short follow up times.

Two other features of the analysis that might be of interest are: (i) Empirical Bayes estimation of variances is considered, an attractive alternative to the standard necessity of choosing between individual estimates and a pooled estimate; (ii) The analysis is done in such a way as to be partially "self-correcting" if some of the individual response curves are actually quadratic over the follow up period.

Section 2 presents the formal model considered, and outlines the plan of attack. Section 3 discusses the needed empirical Bayes theory. Section 4 gives the application to bone loss with age in women. A discussion is given in Section 5.

# 2. MODELING THE PROBLEM

## 2.1 Overview

The procedure proposed in this paper for modeling the problem is based on the following four steps:

1. Fit a straight line to each subject's data points to obtain a least-squares slope estimate of the subject's response curve;
2. Consider each slope estimate to be an estimate of the subject's true response curve slope at a suitable point in the study interval;
3. Assume that, in the population, the individual true slopes are normally distributed with an age-specific population mean;
4. Appropriately model the age-specific population mean.

The justification for this approach is the desire to minimize the structural assumptions in the modeling. Standard approaches tend to require the assumption of specific stochastic models for the individual response curves. Although this may ultimately be needed, the method proposed here is useful at the data analytic stage where good estimates of the individual slopes and population mean may be desired without involving very uncertain modeling structures. The details are presented in the next subsection.

## 2.2 The Model

The response curve of an individual will be denoted $y(t)$, and of interest is its rate of change

$$\beta(t) = \frac{d}{dt}y(t).$$

It will be assumed that, in the population, $\beta(t)$ has a random distribution with mean function $\mu(t)$, i.e.,

$$\beta(t) = \mu(t)+\varepsilon(t), \tag{2.1}$$

where $\varepsilon(t)$ is a random function having mean zero for all t.

Assume the study involves n subjects with measurements $y_{i1}$, . . . , $y_{im_i}$ of the $i^{th}$ subject's response $y_i(t)$ at times $t_{i1} <$ , , , $< t_{im_i}$, i=1, . . ., n. We assume the measurement errors are independent and normally distributed with variance $\sigma_i{}^2$ for the $i^{th}$ individual, i.e.,

$$y_{ij} \sim N(y_i(t_{ij}),\sigma_i{}^2) \qquad j=1, . . ., m_i;\ i=1, . . ., n.$$

Based on these observations, we wish to estimate $\mu(t)$ and the $\beta_i(t) = \frac{d}{dt}y_i(t)$ for the follow-up intervals.

The empirical Bayes or Bayes approach seeks to utilize the information available from (2.1) in the estimation. The difficulties in proceeding directly, however, are considerable, in that the nature of $\varepsilon(t)$ must be carefully specified. Usually, this involves assuming some functional form for $\beta(t)$ (and $\mu(t)$) involving random parameters. Analysis may be very sensitive to the assumed structure, an undesirable feature of this direct approach.

An alternative method of proceeding can be based on the observation that if the follow up interval $(t_{i1}, t_{im_i})$ for the $i^{th}$ individual is short, it is reasonable to expect $y_i(t)$ to be approximately linear on this interval, say

$$y_i(t) \cong \alpha_i+\beta_i t. \tag{2.2}$$

If this approximation is reasonable, the ordinary least squares estimator of $\beta_i$ is given by

$$b_i = \sum_{j=1}^{m_i}(y_{ij}-\overline{y}_i)(t_{ij}-\overline{t}_i)/\sum_{j=1}^{m_i}(t_{ij}-\overline{t}_i)^2 \tag{2.3}$$

where

$$\overline{y}_i = \sum_{j=1}^{m_i} y_{ij}/m_i \quad \text{and} \quad \overline{t}_i = \sum_{j=1}^{m_i} t_{ij}/m_i.$$

This will be (approximately) normally distributed with mean $\beta_i$ and variance

$$d_i = \sigma_i^2/\sum_{j=1}^{m_i}(t_{ij}-\overline{t}_i)^2. \tag{2.4}$$

To estimate the $\sigma_i^2$, we can use the independent random variables

$$s_i^2 = \sum_{j=1}^{m_i}[(y_{ij}-\overline{y}_i)^2-b_i(y_{ij}-\overline{y}_i)(t_{ij}-\overline{t}_i)]\sim\sigma_i^2\chi^2{}_{(m_i-2)}. \qquad (2.5)$$

Although $b_i$ is formally an estimate of $\beta_i(t)$ for all $t\epsilon(t_{i1},t_{im_i})$, it is helpful to consider it an estimate only of $\beta_i(t_i)$, where

$$t_i = \overline{t}_i + \frac{\sum_{j=1}^{m_i}(t_{ij}-\overline{t}_i)^3}{2\sum_{j=1}^{m_i}(t_{ij}-\overline{t}_i)^2}. \qquad (2.6)$$

The reason for this, as shown in Appendix A, is that if $y_i(t)$ is really a quadratic on $(t_{i1},t_{im_i})$, then at $t_i$ the actual slope $\beta_i(t_i)$ will equal the slope found by fitting a linear function to the points $(t_{i1},y_i(t_{i1}))$, . . ., $(t_{im_i},y_i(t_{im_i}))$. The approach thus possesses a degree of built in robustness.

We have reduced the data to independent

$$b_i \sim N(\beta_i(t_i),d_i), \quad i=1, . . ., n. \qquad (2.7)$$

This reduction greatly simplifies the needed modeling of the randomness in (2.1), since we effectively have only one random observation at one time $t_i$ from each realization $\beta_i(t)$. Assuming independence of subjects, progress can now be made solely by modeling the distribution of $\beta(t)$ at fixed points t. We will indeed assume that, for fixed t, the population slopes are distributed as

$$\beta(t) \sim N(\mu(t),D).$$

The assumption of constant (over t) variance D of the distribution of slopes in the population seems reasonable for the example we consider, although more generally a function D(t) might be inserted. The point is that, because of the simple form of the data in (2.7), we need not be concerned with the covariance of $\beta(t)$ at different times, and hence can avoid imposing particular structures on the $\beta_i$ over the entire age span of interest.

It will also be necessary to assume some structure for $\mu(t)$, the age specific population mean rate. The choice of a functional form of $\mu(t)$ is sometimes determined a priori from the theoretical considerations of the underlying biological or physical process. More often, however, no theoretical model exists and $\mu(t)$ is approximated by a polynomial function in t, with the degree of the polynomial determined empirically from the data. Although more general forms of $\mu(t)$ could be considered, we will suppose that $\mu(t)$ can be expressed as

$$\mu(t) = \underset{\sim}{h}(t)\underset{\sim}{\gamma}, \tag{2.8}$$

where $\underset{\sim}{\gamma}$ is a P-vector of regression coefficients and $\underset{\sim}{h}(t) = (h_1(t), . . ., h_p(t))$, where the $h_j(t)$ are given functions of t. In particular, for a polynomial function of degree (P-1), $\underset{\sim}{h}(t) = (1, t, t^2, \ldots, t^{P-1})$.

In our example, it will suffice to consider P=2, $h_1(t)=1$ and $h_2(t)=t$, i.e., a linear $\mu(t)$. To summarize, the problem has been reduced to considering for i=1, . . ., n,

$$b_i \sim N(\beta_i(t_i), d_i), \qquad (2.9)$$

where

$$d_i = \sigma_i^2 / \sum_{j=1}^{m_i} (t_{ij} - \bar{t}_i)^2 \text{ and } s_i^2 \sim \sigma_i^2 \chi^2_{(m_i-2)}, \qquad (2.10)$$

and

$$\beta_i(t_i) \sim N(\underset{\sim}{h}(t_i)\underset{\sim}{\gamma}, D), \qquad (2.11)$$

from which we wish to estimate the $\beta_i(t_i)$ and $\underset{\sim}{\gamma}$.

It should be noted that we have "thrown away" information in the above approach, by compressing the data to the $b_i$ (and the $s_i^2$). In particular, there may be other information in the original $y_{ij}$, or, at least, in the least squares estimate of $\alpha_i$ (see (2.2)). This information cannot be used to estimate the $\beta_i(t)$ and $\mu(t)$, however, unless very special structures for the $\beta_i$ are assumed. The amount of information lost seems marginal, in any case, since the linear approximation to $\beta_i(t)$ on $(t_{i1}, t_{im_i})$ is usually very reasonable and since $\alpha_i$ will tend to be so variable in the population as to provide little information about the $\beta_i(t)$ or $\mu(t)$.

## 3. EMPIRICAL BAYES ANALYSIS

### 3.1 Bayes Versus Empirical Bayes

The problem to be analyzed in (2.9) through (2.11) can be approached from two directions. One is the Bayesian approach of Lindley and Smith (1972), which would involve putting priors (typically noninformative ones) on D and $\underset{\sim}{\gamma}$ and probably a two stage prior on the $\sigma_i^2$, and then calculating the posterior means of the $\beta_i(t_i)$ and of $\underset{\sim}{\gamma}$ given the data.

The second approach is the empirical Bayes approach of calculating Bayes estimates of the desired quantities, pretending that D, $\underset{\sim}{\gamma}$ and the $\sigma_i^2$ are known, and then inserting estimates of D, $\underset{\sim}{\gamma}$ and the $\sigma_i^2$ based on their joint likelihood.

The Bayes approach will usually be superior for small or moderate n, but has the disadvantage of being somewhat remote in the sense that the estimators used must be evaluated by numerical integration and are not always easy to intuitively understand. The empirical Bayes estimators are often of a simpler form, and for large n tend to work as well as the Bayes estimators. Since the n considered in this paper will be large, we will opt for the empirical Bayes approach.

## 3.2 Bayes Estimates of $\beta_i(t_i)$ for Known Parameters

If the $\sigma_i^2$, $\underset{\sim}{\gamma}$, and D are known, then the usual Bayes estimator of $\beta_i(t_i)$ is the posterior mean

$$\hat{b}_i = E(\beta_i(t_i)|b_i) = b_i - \frac{d_i}{D+d_i}(b_i-\mu(t_i)), \qquad (3.1)$$

(recall $\mu(t_i) = \underset{\sim}{h}(t_i)\underset{\sim}{\gamma}$) which has posterior variance

$$\mathrm{Var}\left(\beta_i(t_i)|b_i\right) = \frac{d_i D}{D+d_i}. \qquad (3.2)$$

This is a weighted combination of the least squares estimator $b_i$ and the population mean estimate $\mu(t_i)$. We now must estimate $\underset{\sim}{\gamma}$, D, and the $\sigma_i^2$.

## 3.3 Estimation of Variances -- A Compromise with Pooling

The measurement process (used to obtain the $y_{ij}$) might well be assumed to have equal variance $\sigma^2$ across subjects. On the other hand, the measurements may be more or less variable for certain subjects. Classical analysis usually requires a choice between individual estimates and a pooled estimate. The situation cries out for empirical Bayes (or Bayes) handling, to arrive at a compromise between the two extremes.

Suppose we have

$$s_i^2 \sim \sigma_i^2 \chi^2_{n_i} \qquad i=1, \ldots, k,$$

(let $f_i(s_i^2|\sigma_i^2)$ denote the appropriate Chi-squared density), and believe that the $\sigma_i^2$ are distributed in the population according to a prior distribution $\pi(\sigma^2)$. Empirical Bayes analysis is usually fairly robust with respect to the functional form chosen for $\pi$ (see Berger (1980)), so for convenience we suppose $\pi$ is an inverse gamma distribution with density

$$\pi(\sigma^2) \propto \sigma^{-2(\alpha+1)} e^{-\beta/(2\sigma^2)}.$$

The parameters $\alpha$ and $\beta$ are considered unknown, and are to be estimated from the data. (Again the Bayes approach of using noninformative priors for $\alpha$ and $\beta$ and calculating the posterior means of the $\sigma_i^2$ is a very attractive alternative.)

For known $\alpha$ and $\beta$, an easy calculation shows that the posterior mean for $\sigma_i^2$ is

$$E(\sigma_i^2|s_i^2) = \frac{s_i^2 + \beta}{2\alpha+n_i}. \tag{3.3}$$

To estimate $\alpha$ and $\beta$ from the data, it is common to employ moment methods or maximum likelihood methods. In the interests of providing intuitively accessible formulas, we here use a mixture of these.

The marginal densities

$$m_i(s_i^2|\alpha,\beta) = \int_0^\infty f_i(s_i^2|\sigma_i^2)\pi(\sigma_i^2|\alpha,\beta)d\sigma_i^2,$$

provide the vehicle for the estimation of $\alpha$ and $\beta$. Using the moment approach first, calculation gives

$$\mu_i = E^{m_i}[s_i^2] = E^{\pi(\sigma_i^2|\alpha,\beta)}E^{f_i(s_i^2|\sigma_i^2)}[s_i^2]$$

$$= E^{\pi}[n_i\sigma_i^2] = \frac{n_i\beta}{2(\alpha-1)},$$

and

$$\mathrm{Var}_i = E^{m_i}(s_i^2-\mu_i)^2 = E^{\pi(\sigma_i^2|\alpha,\beta)}E^{f_i(s_i^2|\sigma_i^2)}[s_i^2-\mu_i]^2$$

$$= E^{\pi}E^{f_i}\{(s_i^2-n_i\sigma_i^2)^2+(n_i\sigma_i^2-\mu_i)^2\}$$

$$= E^{\pi}\{2n_i\sigma_i^4+(n_i\sigma_i^2-\mu_i)^2\}$$

$$= \frac{n_i\beta^2}{4(\alpha-1)^2}\left(2+\frac{(2+n_i)}{\alpha-2}\right).$$

Thus the

$$x_i = s_i^2/n_i, \qquad i = 1, \ldots, k,$$

are independent and have marginal means $\beta/[2(\alpha-1)]$ and variances

$$v_i = \frac{\beta^2}{4(\alpha-1)^2 n_i}\left(2+\frac{2+n_i}{\alpha-2}\right).$$

The best linear estimate of the common mean $\beta/[2(\alpha-1)]$ is

$$\hat{s}^2 = \frac{\sum_{i=1}^{k} x_i/v_i}{\sum_{i=1}^{k} 1/v_i} = \frac{\sum_{i=1}^{k} s_i^2/[2(\alpha-1)+n_i]}{\sum_{i=1}^{k} n_i/[2(\alpha-1)+n_i]} .$$

Equating $\hat{s}^2$ and $\beta/[2(\alpha-1)]$, we get as an estimate for $\beta$

$$\frac{2(\alpha-1)\sum_{i=1}^{k} s_i^2/[2(\alpha-1)+n_i]}{\sum_{i=1}^{k} n_i/[2(\alpha-1)+n_i]}.$$

Before proceeding with determination of $\alpha$, it is worthwhile to slightly modify the estimate of $\beta$, by replacing $(\alpha-1)$ with $\alpha$. The change is minor and gives an estimator which reduces in the limits to the usual estimators. Thus we have

$$\hat{\beta} = \frac{2\alpha \sum_{i=1}^{k} s_i^2/[2\alpha+n_i]}{\sum_{i=1}^{k} n_i/[2\alpha+n_i]}. \qquad (3.4)$$

as the estimate of $\beta$.

The method of moments gets very messy in estimating $\alpha$. Hence it seems reasonable to revert back to maximum likelihood. The log likelihood for $\alpha$ and $\beta$ (from $\pi_{i=1}^{k} m_i(s_i^2|\alpha,\beta)$) is (for some constant $K(s_1^2, \ldots, s_k^2)$)

$$\log L(\alpha,\beta) = K+k\ \alpha\ \log\frac{\beta}{2} + \sum_{i=1}^{k} \log \frac{\Gamma(\alpha+\frac{n_i}{2})}{\Gamma(\alpha)} - \sum_{i=1}^{k} (\alpha+\frac{n_i}{2})\log[\frac{1}{2}(s_i^2+\beta)]. \qquad (3.5)$$

Replacing $\beta$ by $\hat{\beta}$ (from (3.4)) and maximizing over $\alpha$, will give an estimate $\hat{\alpha}$ for $\alpha$.

Plugging $\hat{\beta}$ and $\hat{\alpha}$ into (3.3), we get as estimates of $\sigma_i^2$

$$\hat{\sigma}_i^2 = \frac{s_i^2+\hat{\beta}}{2\hat{\alpha}+n_i} = (\frac{n_i}{2\hat{\alpha}+n_i})\frac{s_i^2}{n_i}+(\frac{2\hat{\alpha}}{2\hat{\alpha}+n_i})\frac{\sum_{i=1}^{k} s_i^2/[2\hat{\alpha}+n_i]}{\sum_{i=1}^{k} n_i/[2\hat{\alpha}+n_i]}. \qquad (3.6)$$

Comment 1. If $n_i = n$, $i = 1, \ldots, k$, then

$$\hat{\sigma}_i^2 = (\frac{n}{2\hat{\alpha}+n})\frac{s_i^2}{n}+(\frac{2\hat{\alpha}}{2\hat{\alpha}+n})\frac{\Sigma s_i^2}{\Sigma n_i},$$

which is a compromise between the individual estimators and the pooled estimate.

Comment 2. When $\hat{\alpha} = 0$, $\hat{\sigma}_i{}^2 = \frac{s_i{}^2}{n_i}$ (the individual estimates), while when $\hat{\alpha} = \propto$, $\hat{\sigma}_i{}^2 = \frac{\sum s_i{}^2}{\sum n_i}$ (the pooled estimate).

Comment 3. It can be shown that $\hat{\alpha}$ and hence $\hat{\sigma}_i{}^2$ is scale invariant.

Comment 4. If $s_i{}^2/n_i = \zeta$ for $i = 1, . . ., k$, then logL will be of the form $h(\alpha)+g(\zeta)$, for some functions h and g, and indeed $h(\alpha)$ can be shown to attain a maximum at $\alpha=\infty$. Thus when the $s_i{}^2/n_i$ are nearly equal, $\sigma_i{}^2$ will be the pooled estimate as intuition would demand.

Comment 5. The following comments may be of assistance in maximizing (3.5) over $\alpha$ (with $\beta$ replaced by $\hat{\beta}$):

(i) As can be seen from Comment 4, it is possible for the maximum to be attained at $\alpha=\infty$. It can be shown that $\hat{\alpha}<\infty$ if

$$\Sigma n_i{}^2 - 2\,\Sigma n_i + \frac{(\Sigma n_i)^2(\Sigma s_i{}^4)}{(\Sigma s_i{}^2)^2} - \frac{2(\Sigma n_i)(\Sigma n_i s_i{}^2)}{\Sigma s_i{}^2} > 0.$$

(ii) As $\alpha\to 0$, logL $(\alpha, \hat{\beta}) \to -\infty$, so $\hat{\alpha}$ can never equal zero.

(iii) Note that

$$\frac{d}{d\alpha}\log\frac{\Gamma(\alpha+\frac{n_i}{2})}{\Gamma(\alpha)} = \begin{cases} \sum_{j=1}^{n_i/2}\frac{1}{\alpha+j-1} & \text{if } n_i \text{ is even} \\ \Psi(\alpha+\frac{n_i}{2})-\Psi(\alpha) & \text{if } n_i \text{is odd,} \end{cases}$$

where $\Psi$ is the digamma function. Furthermore,

$$\Psi(\alpha+\frac{n_i}{2})-\Psi(\alpha) = \frac{n_i}{2}\sum_{j=1}^{\infty}\frac{1}{(j+\alpha-1)(j+\alpha-1+\frac{n_i}{2})}.$$

(iv) If $n_i = n$ (even), $i = 1, \ldots, k$,

$$\frac{d}{d\alpha}\log L(\alpha,\hat{\beta}) = \frac{kn}{2}\sum_{j=1}^{n/2}\frac{1}{(\alpha+j-1)} - \sum_{i=1}^{k}\log(1+\frac{kns_i^2}{2\alpha\Sigma s_i^2})-(1+\frac{n}{2\alpha})\sum_{i=1}^{k}(1+\frac{kns_i^2}{2\alpha\Sigma s_i^2})^{-1}.$$

Conclusion: We will be using (3.6) to estimate the $\sigma_i^2$. In our problem, $n_i = m_i-2$, and $k=n$. It should be observed that, again, some information is being ignored, namely the information about the $\sigma_i^2$ contained in (2.9). A grand simultaneous empirical Bayes analysis involving the joint density of everything in sight could have been performed, but the information ignored seems minimal and the advantage for intuitive checking of results by considering the $\sigma_i^2$ separately seems considerable.

## 3.4 Estimation of $\underset{\sim}{\gamma}$ and D.

There are several ways of estimating $\underset{\sim}{\gamma}$ and $\underset{\sim}{D}$. Rubin (1980) gives details of the application of the EM algorithm (Dempster, Laird, and Rubin (1977)) by treating the $\beta_i(t_i)$ as missing data. Other methods are based on the marginal distribution of the $b_i$, which can be seen from (2.9) and (2.11) to be given by

$$b_i \sim N(\underset{\sim}{h}(t_i)\underset{\sim}{\gamma}, D+d_i),\ i=1, \ldots, n. \tag{3.7}$$

The moment (or weighted least squares) procedure (see Fay and Herriot (1979)) and the maximum likelihood procedure (see Efron and Morris (1975)) are two such methods. The maximum likelihood procedure is simple here, and leads to reasonably intuitive estimators, so we employ this method.

The likelihood function L for $\underset{\sim}{\gamma}$ and D is just the product of the normal densities in (3.7). It is easy to check that the maximizing value of $\underset{\sim}{\gamma}$ is

$$\underset{\sim}{\hat{\gamma}} = (\underset{\sim}{t}'\underset{\sim}{V}^{-1}\underset{\sim}{t})^{-1}(\underset{\sim}{t}'\underset{\sim}{V}^{-1}\underset{\sim}{b}), \tag{3.8}$$

where

$$\underset{\sim}{t}' = (t_1, \ldots, t_i, \ldots, t_n)$$

$$\underset{\sim}{V} = D\underset{\sim}{I}_n+\text{diag}\{d_i\},$$

and

$$\underset{\sim}{b} = (b_1, \ldots, b_i, \ldots, b_n)'.$$

Finally, setting $\partial L/\partial D = 0$ results in the equation for D

$$D = \frac{\sum_{i=1}^{n} [\{(b_i - \underset{\sim}{h}(t_i)\underset{\sim}{\gamma})^2 - d_i\}/(D+d_i)^2]}{\sum_{i=1}^{n} (D+d_i)^{-2}} . \qquad (3.9).$$

Equations (3.8) and (3.9) can be solved iteratively for D and $\underset{\sim}{\gamma}$, convergence to a solution generally being very rapid. (To check the uniqueness of the solution, it is a fairly easy task to insert (3.8) into the likelihood function and roughly graph it as a function of D.)

It is interesting to note that (3.8) implies that the estimated $\mu(t)$ will be the weighted least squares estimate based on $\underset{\sim}{b}$ (with the $(D+d_i)^{-1}$, the inverses of the marginal variances of the $b_i$, as weights). Also, if the weighted least squares estimate of $\mu(t)$ is calculated based on the empirical Bayes estimates $\hat{b}_i$ (where now the weights are $(D+d_i)/D^2$, the inverses of the marginal variances of $\hat{b}_i$) one gets the same result.

If n were small, it would be better to use the modifications of the above estimate suggested by Morris (1982). For the large n we consider, however, these modifications are not needed.

# 4. APPLICATION TO BONE LOSS IN POSTMENOPAUSAL WOMEN

## 4.1 Background

Osteoporosis, a condition of diminished bone mass with increased risk of fractures, is often considered an accelerated aging process rather than a distinct disease. This problem is most prevalent in postmenopausal Caucasian women. Hence, knowledge about the natural history of bone loss in a general population of these women may improve the understanding of osteoporosis. To this end, a longitudinal study was started in 1971 to characterize the change of bone mass with age in postmenopausal women.

The details of the study design and methodology have been reported previously by Smith et al. (1975). Briefly, all the subjects were volunteers free of diseases known to affect bone metabolism. New subjects have been added to the study whenever they became available. Since the subjects were not compensated for their participation in the study, visits were always scheduled at their convenience and bone mass measurements were made at each visit.

The analysis in this paper is aimed at characterizing the change of bone mass with age in 268 postmenopausal Caucasian women. Their initial ages range from 50 to 95 and the number of visits per subject varies between 3 and 44.

## 4.2 Results

Since all subjects were aged 50 and over, and rapid bone loss is usually assumed to start at around age 50, a convenient transformation, $t$ = age-50 was made. Examination of a typical set of data (see Figure 1) indicated that linearity of the response curve over the follow-up period is a reasonable assumption. Thus for each individual, the least squares slopes $b_i$ (see (2.3)), the sample variances $s_i^2$ (see (2.5)), and the $t_i$ (see (2.6)) were calculated. Since, marginally, the $b_i$ would be distributed $N(\mu(t_i), D+d_i)$, a feeling for the shape of $\mu(t)$ can be found by plotting $b_i$ against $t_i$. This plot is given in Figure 2 (with one apparent outlier off the graph) and suggests that $\mu(t)$ can be approximated by a linear function $\gamma_1+\gamma_2 t$. Thus in (2.8), $\underset{\wedge}{\gamma} = (\gamma_1,\gamma_2)'$ and $\underset{\wedge}{h}(t) = (1,t)$ was used.

In estimating the $\sigma_i^2$ using the technique of subsection 3.3, the likelihood function (3.5) attained a maximum at $\alpha = \infty$, i.e., the empirical Bayes estimate (3.6) turned out to be simply the pooled estimate of variance

$$\hat{\sigma}^2 = \frac{\sum s_i^2}{\sum (m_i-2)} = 1.089\times 10^{-5}.$$

(This was a comforting result, since the measurement process was such that a pooled estimate of variance seemed natural.) This estimate was used in (2.4) and gave estimated variances $d_i$ (of the least squares slopes $b_i$) ranging from $8.7\times10^{-6}$ to 0.15. Finally, equations (3.8) and (3.9) were solved iteratively, yielding estimates

$$\hat{D} = 6.958\times10^{-5}$$

and

$$(\hat{\gamma}_1,\hat{\gamma}_2) = (-0.01754, 4.652\times10^{-4}).$$

The standard errors of these estimates (from the weighted least squares regression based on (3.7) and assuming D and the $d_i$ known) are $1.32\times10^{-3}$ and $5.45\times10^{-5}$. These are probably underestimates of the error, since D and the $d_i$ were estimated. The fitted line for $\mu(t)$ is indicated by the solid line in Figure 2. (As a check, a quadratic was also fitted to $\mu(t)$, but the quadratic term was found to be negligible and statistically insignificant.)

As a check on the modeling assumptions made, a normal probability plot of the standardized residuals

$$r_i = \frac{1}{\sqrt{\hat{D}+d_i}}[b_i-(\hat{\gamma}_1+\hat{\gamma}_2 t_i)]$$

is given in Figure 3 (with two $r_i \sim -4$ excluded). (By (3.7), the $r_i$ should be N(0,1), ignoring the randomness introduced by the estimation of the parameters.) As can be seen, the normality assumptions seem reasonable.

The estimates of D, $d_i$, and $\mu(t_i)$ were then used in (3.1), resulting in empirical Bayes estimates $\hat{b}_i$ of the individual slopes $\beta_i(t_i)$. These estimates are plotted against $t_i$ in Figure 4, and are, as expected, pulled in towards the estimated $\mu(t)$.

## 5. DISCUSSION

Comment 1. The empirical Bayes methods discussed in this paper are recommended as a practical tool in the analysis of longitudinal studies with irregular follow-up intervals, particularly in early stages of the study when most of the follow-up intervals are short. The methods provide an estimate of the population growth rate without requiring specific knowledge concerning the form of the individual growth curves, and lead to what is probably a substantially more accurate picture of the variation of the individual growth rates.

The ultimate medical goal is, of course, to come up with good diagnostics. Improved estimation of the individual growth rates after a short follow-up period is certainly helpful here, but what is really desired is ability to predict the long term future growth rate of a patient. This necessarily involves considerably more involved modeling of the individual growth curves, and probably a fairly substantial sample with long follow-up periods.

Comment 2. Since $\mu(t)$ appears to be approximately linear, one would expect the individual growth curves $y_i(t)$ to be approximately quadratic in the long term. The use of straight line fits to subjects with long follow-up intervals thus seems suspect. However, the use of $t_i$ (see (2.6)) as the point in the follow-up interval at which the slope is considered to be estimated theoretically protects against quadratic growth curves.

To investigate the success of this method, a quadratic function in t was fitted to the measurements of each of the women in the study. Eleven of these functions had a significant quadratic term. If the estimated regression function is

$$\hat{y}_i(t) = b_{i0}+b_{i1}t+b_{i2}t^2,$$

a reasonable estimate of the actual slope at $t_i$ is

$$\hat{y}_i'(t_i) = b_{i1}+2b_{i2}t_i.$$

The absolute differences between $\hat{y}_i'(t_i)$ and $b_i$ for the eleven women were all less than 0.66 times the estimated standard deviation, $\sqrt{d_i}$, of the $b_i$, (average absolute difference = $0.11\sqrt{d_i}$) so the use of $t_i$ seems to essentially eliminate the problem of possible quadratic growth curves. (A true underlying quadratic growth curve would tend to cause an inflated variance $d_i$ for $b_i$, but a few somewhat enlarged variances should not affect things too seriously.)

Comment 3. The estimated population growth rate

$$\hat{\mu}(t) = -0.01754+4.652\times10^{-4}(t-50)$$

has the surprising feature that it predicts an increase in bone mass after age 87. This is near the end of the study period, so the phenomenon could well be an artifact of the assumed linearity of $\mu(t)$. However, examination of the individual least squares rates of women over age 70 with five or more years of follow-up revealed that about one-third of them did have significantly (at 5% level) positive rates, indicating the existence of positive rates for at least a substantial proportion of women over 70.

An increase in bone mass is biologically plausible because while some old bone is absorbed on the inside of the long bones, new bone is deposited on the outside. Thus $\beta_i(t)$ can be considered to be a sum of two rates, one positive and one negative, and the positive rate could certainly dominate after a certain age. Even if there is a net gain in bone mass after a certain age, however, it is not clear that there is an increase in the mechanical strength of the bone and hence a decrease in the risk of fracture. These questions are now being investigated.

## APPENDIX A. JUSTIFICATION FOR (2.6)

Suppose $y_i(t)$ is really a quadratic function

$$y_i(t) = b_{i0}+b_{i1}t+b_{i2}t^2, \tag{A.1}$$

so that

$$\beta_i(t) = \frac{d}{dt}\, y_i(t) = b_{i1}+2b_{i2}t. \tag{A.2}$$

Imagine that we fit a linear function (a+bt) to $y_i(t)$, based on observing $y_i(t)$ exactly at the points $t_{i1}, \ldots, t_{im_i}$. The least squares value for the slope b is

$$b = \frac{\sum_{j=1}^{m_i} y_i(t_{ij})t_{ij} - \frac{1}{m_i}[\sum_j y_i(t_{ij})][\sum_j t_{ij}]}{\sum_j t_{ij}^2 - \frac{1}{m_i}[\sum_j t_{ij}]^2}$$

$$= b_{i1}+2b_{i2}t_i$$

using (A.1), (2.6), and some simplification. Thus, from (A.2), b is precisely equal to $\beta_i(t_i)$. Of course, in reality, the observations are not the $y_i(t_{ij})$ exactly (but rather the $y_i(t_{ij})$ plus error), so the correspondence will not be exact.

## Figure Captions

Figure 1: Schematic diagram of measurements in a longitudinal study with irregular follow-up intervals. Consecutive measurements of each subject are joined by straight lines.

Figure 2: Individual least squares estimates of rate of bone loss $b_i$ vs. ti, where the ti are suitably chosen points in the follow-up intervals.

Figure 3: Normal probability plot of standardized residuals.

Figure 4: Individual empirical Bayes estimates of rate of bone loss $\hat{b}_i$ vs. $t_i$.

# REFERENCES

Berger, J. (1980). Statistical Decision Theory: Foundations, Concepts, and Methods. Springer-Verlag, New York.

Carter, Grace M., and Rolph, John E. (1974), "Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities," Journal of the American Statistical Association, 69, 880-885.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Ser. B, 39, 1-38.

Efron, Bradley, and Morris, Carl (1973), "Stein's Estimation Rule and Its Competitors -- An Empirical Bayes Approach," Journal of the American Statistical Association, 68, 117-130.

Efron, Bradley, and Morris, Carl (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," Journal of the American Statistical Association, 70, 311-319.

Fay, Robert E. III, and Heriot, Roger A. (1979), "Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data," Journal of the American Statistical Association, 74, 269-277.

Fearn, T. (1975), "A Bayesian Approach to Growth Curves," Biometrika, 62, 89-100.

Grizzle, James E., and Allen, David M. (1969), "Analysis of Growth and Response Curves," Biometrics, 25, 357-381.

James, W., and Stein, C.M. (1961), "Estimation with Quadratic Loss," in Proc. 4th Berkeley Symposium, 1, 361-379.

Lindley, D.V., and Smith, A.F.M. (1972), "Bayes Estimates for the Linear Model," (with discussion), Journal of the Royal Statistical Society, Ser. B, 34, 1-41.

Morris, C. (1982), "Parametric Empirical Bayes Inference: Theory and Applications," Invited paper for the Journal of the American Satsitical Association.

Pothoff, R.F., and Roy, S.N. (1964), "A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems," Biometrika, 51, 313-326.

Rao, C.R. (1973), Linear Statistical Inference and Its Applications, New York: John Wiley & Sons.

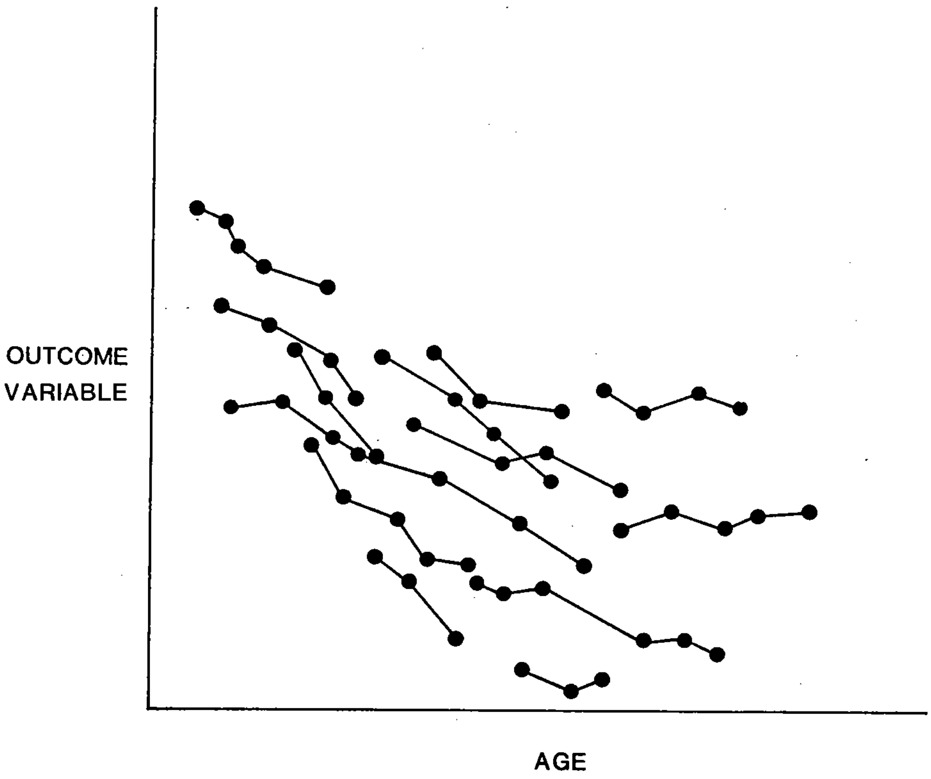
Rolph, John E., Williams, Alert P., and Lee, Carolyn L. (1979), "Medical School Admission and Residence of Applicants: Empirical Bayes Estimates of Logit Coefficients," Journal of Educational Statistics, 4, 291-323.

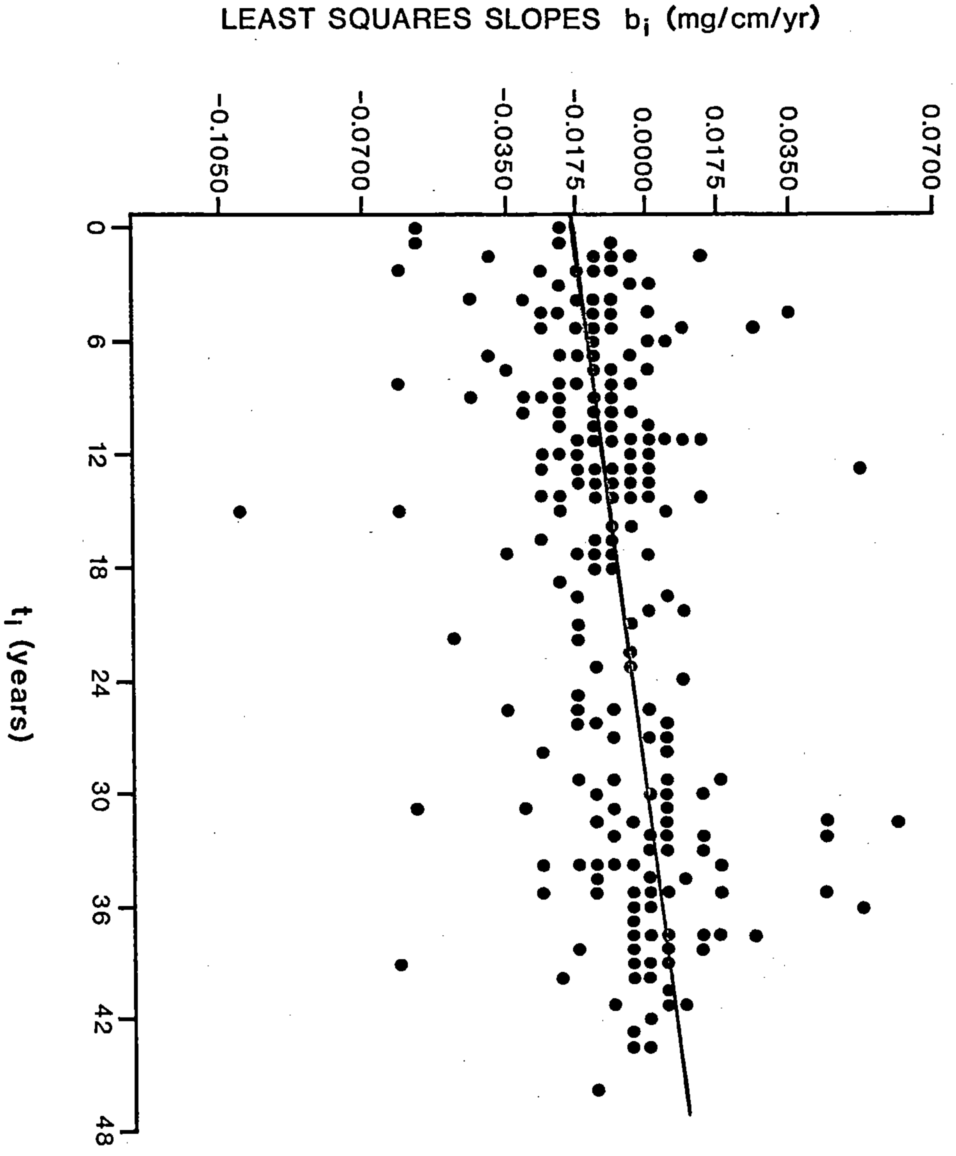
Rubin, Donald B., (1980), "Using Empirical Bayes Techniques in the Law School Validity Studies," Journal of the American Statistical Association, 372, 801-816.

Smith, D.M., Khairi, M.R.A., and Johnston, C.C., Jr. (1975), "The Loss of Bone Mineral with Aging and Its Relationship to Risk of Fracture," Journal of Clinical Investigations, 56, 311-318.

Zerbe, Gary O., and Walker, Strother H. (1977), "A Randomization Test for Comparison of Groups fo Growth Curves with Different Polynomial Design Matrices," Biometrics, 33, 653-657.

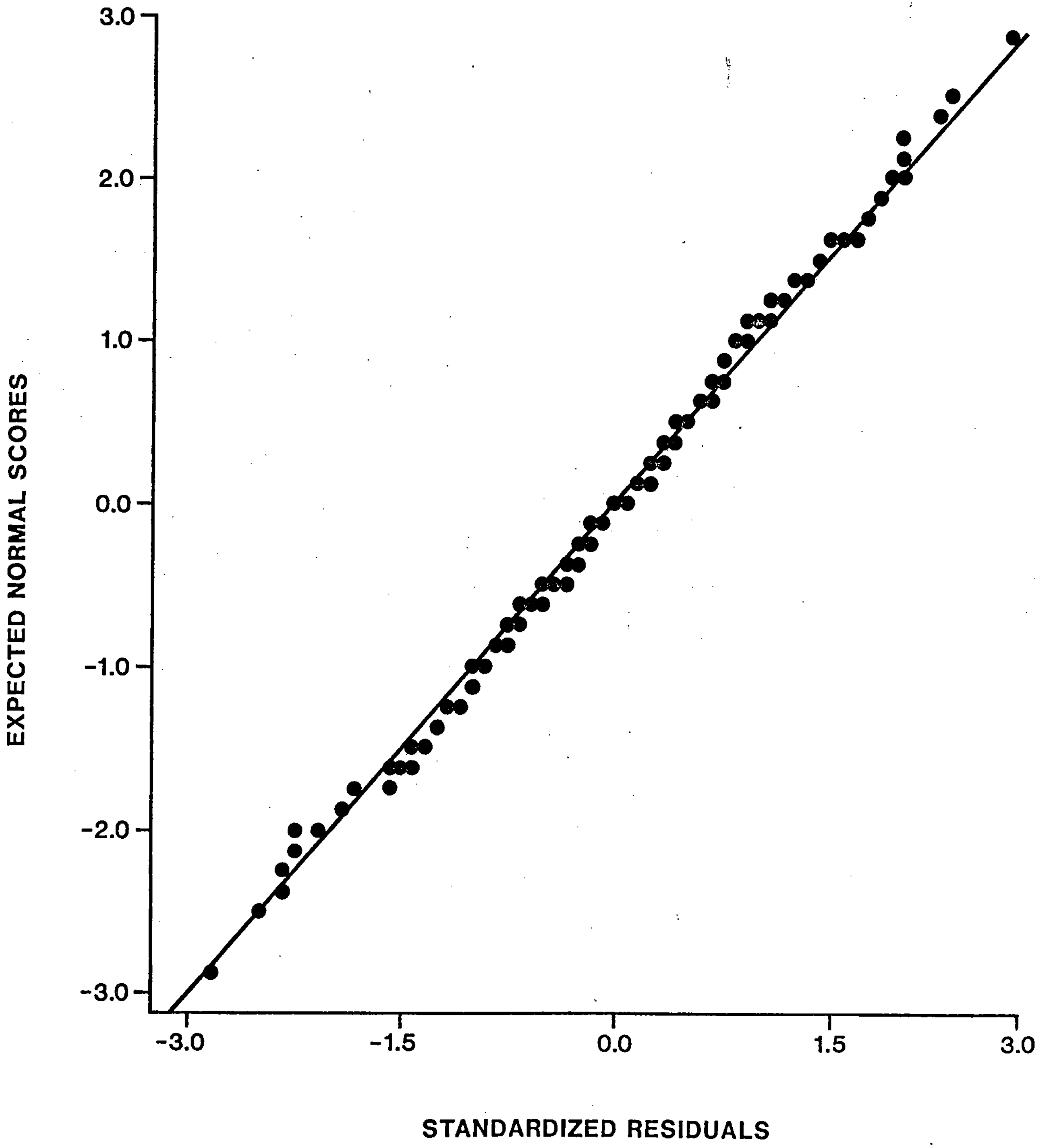
Fig. 1





LEAST SQUARES SLOPES $b_i$ (mg/cm/yr)
-0.1050
-0.0700
-0.0350
-0.0175
0.0000
0.0175
0.0350
0.0700
0
6
12
18
24
30
36
42
48
$t_i$ (years)

Fig. 2



3.0
2.0
1.0
0.0
-1.0
-2.0
-3.0
EXPECTED NORMAL SCORES
-3.0
-1.5
0.0
1.5
3.0
STANDARDIZED RESIDUALS

Fig. 3

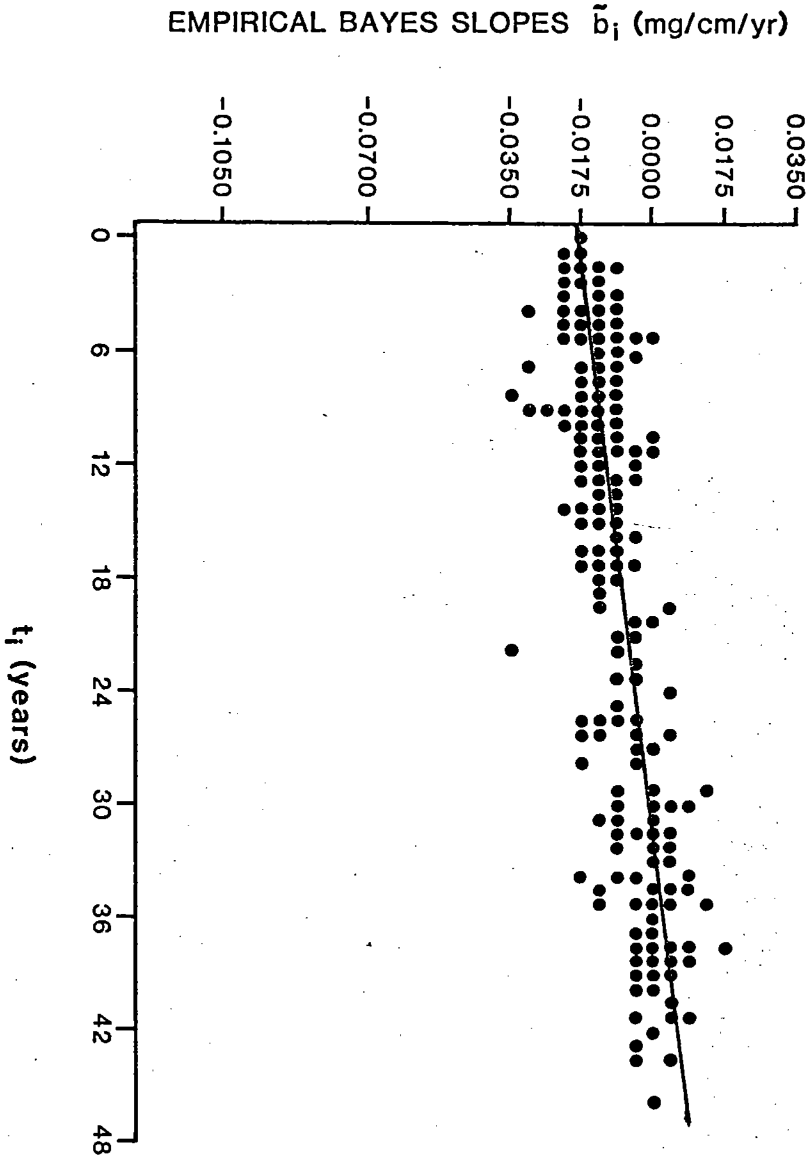

EMPIRICAL BAYES SLOPES $\tilde{b}_i$ (mg/cm/yr)
0.0350
0.0175
0.0000
-0.0175
-0.0350
-0.0700
-0.1050
0
6
12
18
24
30
36
42
48
$t_i$ (years)

Fig. 4