Statistical Science
1987, Vol. 2, No. 4, 434–471

Prediction of Future Observations in Growth Curve Models

C. Radhakrishna Rao

Abstract. The problem of predicting a future measurement on an individual given the past measurements is discussed under nonparametric and parametric growth models. The efficiencies of different methods of prediction are assessed by cross-validation or leave-one-out technique in each of three data sets and the results are compared. Under nonparametric models, direct and inverse regression methods of prediction are described and their relative advantages and disadvantages are discussed. Under parametric models polynomial and factor analytic type growth curves are considered. Bayesian and empirical Bayesian methods are used to deal with unknown parameters. A general finding is that much of the information for forecasting is contained in the immediate past few observations or a few summary statistics based on past data. A number of data reduction methods are suggested and analyses based on them are described. The usefulness of the leave-one-out technique in model selection is demonstrated. A new method of calibration is introduced to improve prediction.

Key words and phrases: Bayesian approach, calibration, cross-validation, empirical Bayes, factor analytic model, inverse regression, leave-one-out method, mixed model, part correlation, polynomial model, predictive density, principal component regression.

1. INTRODUCTION

Let Y_i denote a measurement taken on an individual at time t_i, $i = 1, 2, \cdots$. A problem of great interest is the prediction of $Y_{p+1}, Y_{p+2}, \cdots$ having observed the values of $Y_1 = y_1, \cdots, Y_p = y_p$ on an individual at p previous time points $t_1 < \cdots < t_p$. Generally, we have previously recorded data on all the measurements $Y_1, \cdots, Y_p, Y_{p+1}, \cdots$ taken on, say, n individuals from a related or the same population to which the individual for whom prediction is required belongs. What is the best way of using the information in the recorded data for constructing a prediction formula for future observations on new individuals? The data on the past individuals and the current (or future) individual for whom prediction is required may be represented as in Table 1.

There is considerable literature on this problem. We briefly review some of the known results and discuss some alternative approaches. We choose for illustration three data sets: weights of 13 mice taken at seven time points (Table 2), ramus heights (the ramus is the ascending part of the mandible) of 20 boys taken at four different ages (Table 3) and dental measurements of 11 girls and 16 boys taken at four different ages (Table 4). In each case we use the previous measurements to predict the last measurement and compare with the actual observed value to assess the accuracy of any given method.

2. GENERAL THEORY

To simplify the notation let us represent by vectors U and W the two sets of variables (observed and to be predicted),

$$U = (Y_1, \cdots, Y_p)', \quad W = (Y_{p+1}, Y_{p+2}, \cdots)'. \tag{2.1}$$

Then the complete data on the past n individuals can be represented as

$$(U_i' : W_i') = (y_{1i}, \cdots, y_{pi}, y_{p+1,i}, y_{p+2,i}, \cdots), \quad i = 1, \cdots, n, \tag{2.2}$$

and the measurements on the current individual by

$$(U_c' : W_c') \tag{2.3}$$

C. Radhakrishna Rao is University Professor at the University of Pittsburgh and National Professor in India. His mailing address is: Department of Mathematics and Statistics, Thackeray Hall, University of Pittsburgh, Pittsburgh, Pennsylvania 15260.

TABLE 1
Measurements at different time points

Individuals	Time Points				
	t_1	$\cdots$	t_p	t_{p+1}	t_{p+2}
Past					
1	y_{11}	$\cdots$	y_{p1}	$y_{p+1,1}$	$y_{p+2,1}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
n	y_{1n}	$\cdots$	y_{pn}	$y_{p+1,n}$	$y_{p+2,n}$
Current					
c	y_{1c}	$\cdots$	y_{pc}	?	?

The values to be predicted are indicated by ?

TABLE 2
Weights of 13 *male mice measured at successive intervals of* 3 *days over* 21 *days from birth to weaning* (*Williams and Izenman*, 1981)

	Day 3	Day 6	Day 9	Day 12	Day 15	Day 18	Day 21
1	0.190	0.388	0.621	0.823	1.078	1.132	1.191
2	0.218	0.393	0.568	0.729	0.839	0.852	1.004
3	0.211	0.394	0.549	0.700	0.783	0.870	0.925
4	0.209	0.419	0.645	0.850	1.001	1.026	1.069
5	0.193	0.362	0.520	0.530	0.641	0.640[a]	0.751
6	0.201	0.361	0.502	0.530	0.657	0.762	0.888
7	0.202	0.370	0.498	0.650	0.795	0.858	0.910
8	0.190	0.350	0.510	0.666	0.819	0.879	0.929
9	0.219	0.399	0.578	0.699	0.709	0.822	0.953
10	0.225	0.400	0.545	0.690	0.796	0.825	0.836
11	0.224	0.381	0.577	0.756	0.869	0.929	0.999
12	0.187	0.329	0.441	0.525	0.589	0.621	0.796
13	0.278	0.471	0.606	0.770	0.888	1.001	1.105

[a] This could be a recording error, but no change was made in the present computations.

TABLE 3
Ramus heights of 20 *boys at different ages* (*Elston and Grizzle*, 1962; *Grizzle and Allen*, 1969; *Lee and Geisser*, 1975)

	8 yr	8½ yr	9 yr	9½ yr
1	47.8	48.8	49.0	49.7
2	46.4	47.3	47.7	48.4
3	46.3	46.8	47.8	48.5
4	45.1	45.3	46.1	47.2
5	47.6	48.5	48.9	49.3
6	52.5	53.2	53.3	53.7
7	51.2	53.0	54.3	54.5
8	49.8	50.0	50.3	52.7
9	48.1	50.8	52.3	54.4
10	45.0	47.0	47.3	48.3
11	51.2	51.4	51.6	51.9
12	48.5	49.2	53.0	55.5
13	52.1	52.8	53.7	55.0
14	48.2	48.9	49.3	49.8
15	49.6	50.4	51.2	51.8
16	50.7	51.7	52.7	53.3
17	47.2	47.7	48.4	49.5
18	53.3	54.6	55.1	55.3
19	46.2	47.5	48.1	48.4
20	46.3	47.6	51.3	51.8

TABLE 4
Dental measurements of 11 *girls and* 16 *boys* (*Potthoff and Roy*, 1964; *Lee and Geisser*, 1975)

	8 yr	10 yr	12 yr	14 yr
Girls				
1	21	20	21.5	23
2	21	21.5	24	25.5
3	20.5	24	24.5	26
4	23.5	24.5	25	26.5
5	21.5	23	22.5	23.5
6	20	21	21	22.5
7	21.5	22.5	23	25
8	23	23	23.5	24
9	20	21	22	21.5
10	16.5	19	19	19.5
11	24.5	25	28	28
Boys				
12	26	25	29	31
13	21.5	22.5	23	26.5
14	23	22.5	24	27.5
15	25.5	27.5	26.5	27
16	20	23.5	22.5	26
17	24.5	25.5	27	28.5
18	22	22	24.5	26.5
19	24	21.5	24.5	25.5
20	23	20.5	31	26
21	27.5	28	31	31.5
22	23	23	23.5	25
23	21.5	23.5	24	28
24	17	24.5	26	29.5
25	22.5	25.5	25.5	26
26	23	24.5	26	30
27	22	21.5	23.5	25

or simply by $(U' : W')$ dropping the suffix c. We use both the notations depending on the context. The problem we discuss is the prediction of W_c (or W) for the current individual having observed U_c (or U).

We assume that the measurements (U, W) on an individual are the realizations of a stochastic process depending on two vector parameters β and δ. The parameter β is specific to an individual (depending say on a genetic factor) and may vary over the individuals of a given population. The parameter δ is global in character (representing say an environmental factor), common to all individuals of the population. We represent the probability density of (U, W) for an individual with given β and specified δ by $D(U, W \mid \beta, \delta)$ and that of β of the individuals of the population by $D(\beta \mid \theta)$ depending on a vector parameter θ. ($D(\beta \mid \theta)$ is in the nature of a prior density in a Bayesian approach except that, in our context, it has reference to the particular population from which the individuals are sampled. A further prior on δ and θ would make the set-up completely Bayesian. See, for instance, Lee and Geisser (1972).

The joint density of the hypothetical (unobservable) variable β, the predictor variable U and the predictand

W for given θ and δ can be factorized in different ways to provide conditional or predictive distributions (pred's) for predicting β and/or W in terms of observed U and specified θ and δ. In the following formulas, the suffix "obs" refers to the density of observed variables and "pred" to those to be predicted. The quantities on the righthand side of the partition indicate conditioning variables and parameters.

$$\begin{aligned} & D(\beta, U, W \mid \theta, \delta) \\ & \quad = D(\beta \mid \theta) D(U, W \mid \beta, \delta) \\ \text{(2.4)} \quad & \quad = D(\beta \mid \theta) D_{\text{obs}}(U \mid \beta, \delta) \\ & \qquad \cdot D_{\text{pred}}(W \mid U, \beta, \delta) \\ \text{(2.5)} \quad & \quad = D_{\text{obs}}(U \mid \theta, \delta) D_{\text{pred}}(\beta, W \mid U, \theta, \delta) \\ \text{(2.6)} \quad & \quad = D_{\text{obs}}(U \mid \theta, \delta) D_{\text{pred}}(\beta \mid U, \theta, \delta) \\ & \qquad \cdot D_{\text{pred}}(W \mid U, \beta, \delta) \\ \text{(2.7)} \quad & \quad = D_{\text{obs}}(U \mid \theta, \delta) D_{\text{pred}}(W \mid U, \theta, \delta) \\ & \qquad \cdot D_{\text{pred}}(\beta \mid U, W, \theta, \delta). \end{aligned}$$

The preds in (2.4) to (2.7) could be used depending on the nature of the information available on the parameters θ, δ to predict β and W. In this paper, we examine the prediction of W only and note that the prediction of the hypothetical β, which is important in some genetic studies (see Rao, 1953), can be done in a similar way.

Let U_c be the observation on U for the current individual and W_c the observation on W to be predicted. If θ and δ are known, then the appropriate pred is

$$\text{(2.8)} \qquad D_{\text{pred}}(W_c \mid U_c, \theta, \delta)$$

as in (2.7), from which an appropriate predictor $f(U_c)$ of W_c may be obtained to minimize the expected value of a given loss function $L[W_c, f(U_c)]$, when the expectation is taken over future individuals drawn from the population under consideration, i.e. a population characterized by a specified $D(\beta \mid \theta)$ and a specified $D(U, W \mid \beta, \delta)$. For instance, if

$$L[W_c, f(U_c)] = (W_c - f(U_c))^2$$

then

$$\begin{aligned} & f(U_c) = f(U_c, \theta, \delta) \\ \text{(2.9)} \quad & \quad = \int W_c \, D_{\text{pred}}(W_c \mid U_c, \theta, \delta) \, dW_c. \end{aligned}$$

If θ and δ are not known we have several possibilities.

2.1 BAYPRED (Bayesian Prediction)

We choose a prior density function $D_a(\theta, \delta)$ for the unknown (θ, δ) and compute the posterior density function $D_{\text{post}}(\theta, \delta)$ based on available data (see Table 1),

$$\begin{aligned} & D_{\text{post}}(\theta, \delta) \\ \text{(2.1.1)} \quad & \quad = D_{\text{post}}(\theta, \delta \mid U_i, W_i, i = 1, \cdots, n, \text{ and } U_c) \\ & \quad = \frac{D_a(\theta, \delta) L(\theta, \delta)}{\int D_a(\theta, \delta) l(\theta, \delta) \, d\theta \, d\delta}, \end{aligned}$$

where

$$l(\theta, \delta) = D_{\text{obs}}(U_c \mid \theta, \delta) \prod_{i=1}^{n} D_{\text{obs}}(U_i, W_i \mid \theta, \delta)$$

is the likelihood of (θ, δ) given all the observed data. Then the BAYPRED of W_c given U_c is

$$\begin{aligned} & D_{\text{Baypred}}(W_c \mid U_c) \\ \text{(2.1.2)} \quad & \quad = \int D_{\text{pred}}(W_c \mid U_c, \theta, \delta) D_{\text{post}}(\theta, \delta) \, d\theta \, d\delta, \end{aligned}$$

which depends only on the known U_c. The predictor derived from (2.1.2) instead of (2.8) minimizes the loss function in a superpopulation of individuals determined by the chosen a priori density for (θ, δ).

Note. In deriving the posterior distribution of (θ, δ), the current observation U_c is also used. There may be an advantage in not using U_c especially as it enables the derivation of the prediction function applicable to all future observations on U.

2.2 EMPRED (Empirical Prediction)

In this procedure, instead of using the posterior distribution (2.1.1) of (θ, δ) which depends on the chosen a priori distribution, point estimates of θ and δ are obtained from past data and substituted for true values in the appropriate predictive distribution. For instance, if $\hat{\theta}$ and $\hat{\delta}$ are maximum likelihood estimates (mle) of θ and δ obtained by maximizing the likelihood

$$\text{(2.2.1)} \quad D_{\text{obs}}(U_c \mid \theta, \delta) \prod_{i=1}^{n} D_{\text{obs}}(U_i, W_i \mid \theta, \delta)$$

the estimated pred called EMPRED (empirical predictive density) is

$$\text{(2.2.2)} \qquad D_{\text{empred}}(W_c \mid U_c, \hat{\theta}, \hat{\delta}).$$

In using (2.2.2) for predicting W_c, we behave as if $(\hat{\theta}, \hat{\delta})$ is the actual value of the unknown (θ, δ), so that the accuracy of our prediction depends on how close $(\hat{\theta}, \hat{\delta})$ is to the true value (θ, δ). There seems to be no appropriate theory for taking the estimation errors in (θ, δ) into consideration especially when the same estimate $(\hat{\theta}, \hat{\delta})$ is used repeatedly in predicting W for future individuals. We shall see later on that it is possible to make an assessment of the loss involved in using any particular estimate repeatedly in future predictions which may be of help in making an optimum choice of an estimate.

It may be of interest to ask whether W_c for the current individual could be predicted using U_c only and the known form of an individual's stochastic growth process depending on the individual's own parameter β. For this, we need the pred

$$D_{\text{pred}}(W_c \mid U_c, \beta, \delta) \tag{2.2.3}$$

as in (2.4). Because β is unknown, we need an estimate such as that obtained by maximizing

$$D_{\text{obs}}(U_c \mid \beta, \hat{\delta}), \tag{2.2.4}$$

where $\hat{\delta}$ is the estimate of δ from (2.2.1). If $\hat{\beta}$ is the estimate so derived, then the EMPRED for individual prediction is

$$D_{\text{empred}}(W_c \mid U_c, \hat{\beta}, \hat{\delta}). \tag{2.2.5}$$

The prediction obtained from the EMPRED in (2.2.5) does not depend on $D(\beta \mid \theta)$ and, in general, is different from that obtained from the EMPRED in (2.2.2). The relative advantages and disadvantages of these estimates are well known (see Rao, 1975). In our present investigation, we will be comparing the estimates obtained from the EMPRED's (2.2.2) and (2.2.5).

A Bayesian approach is to consider a prior distribution of β and take the expectation of (2.2.3) over β.

2.3 Point Predictor with a Model Choice

Let $h(U)$ be a predictor of a future observation W given U, the previous observations on an individual, and $L(W, h(U))$ be the loss incurred, where W is the observed and $h(U)$ is its predicted value. Then the optimum choice of h is one which minimizes

$$E[L(W, h)], \tag{2.3.1}$$

where the expectation is taken over the PRED, $D_p(W \mid U, \theta, \delta)$ as defined in (2.6). For instance, when W is a scalar variable, the optimum h is the mean, median and the mode of the PRED for the squared error, absolute error and zero-one loss functions, respectively. Let $h_*(U, \theta, \delta)$ be the optimum choice of h in (2.3.1) for given (θ, δ). If (θ, δ) is not known, a natural choice of the predictor is $h_*(U, \hat{\theta}, \hat{\delta})$ where $(\hat{\theta}, \hat{\delta})$ is an estimate of (θ, δ) obtained from past records. In practice we will be interested in assessing the loss incurred in using particular estimates $(\hat{\theta}, \hat{\delta})$ for predicting W given U on future individuals drawn from the specified population. This loss, for chosen $(\hat{\theta}, \hat{\delta})$, is

$$R(\theta, \delta; \hat{\theta}, \hat{\delta}) = E[L(W, h_*(U, \hat{\theta}, \hat{\delta}))], \tag{2.3.2}$$

where the expectation is taken with respect to the joint density function $D_0(U, W \mid \theta, \delta)$ of (U, W) as defined in (2.7). If we consider repetitions of past data, the overall loss can be obtained by taking a further expectation of (2.3.2) over the distribution of $(\hat{\theta}, \hat{\delta})$. In fact such an expectation may provide a criterion for choosing an appropriate estimator of (θ, δ) from past data. For instance, if the mle of (θ, δ) is considered, one could investigate whether there is an asymptotic gain in correcting the mle for bias up to terms of order (n^{-1}). Such a possibility exists as shown by Cox (1975) while investigating a similar problem. (The methods for correcting mle for possible bias are discussed in Bartlett (1953) and Rao (1963).)

In practice, the actual functional forms of the probability densities may not be known. In such a case we may pose the problem as one of making an appropriate choice of the predictor $h(U, \hat{\theta}, \hat{\delta})$ from among a given class of functions H and a given class of estimators S of (θ, δ) from past data. We propose a feasible solution to this problem through the LOO (leave-one-out) technique as described by Lachenbruch (1975) or CV (cross-validation) as discussed in papers by Geisser (1975a) and Stone (1974, 1977).

Let $h(U, \theta, \delta)$ be a chosen functional form and $(\hat{\theta}_{(-i)}, \hat{\delta}_{(-i)})$ be an estimate of (θ, δ) obtained by minimizing

$$\sum_{j=1}^{i-1} L(W_j, h(U_j, \theta, \delta)) + \sum_{j=i+1}^{n} L(W_j, h(U_j, \theta, \delta)), \tag{2.3.3}$$

where L is a given loss function and (U_i, W_i), $i = 1, \cdots, n$, are the past observations on the complete set (U, W). (Note that the observation (U_i, W_i) does not occur in (2.3.2).) Then we define the CVAE (cross-validation assessment error)

$$\text{CVAE}(h) = n^{-1} \sum_{i=1}^{n} L(W_i, h(U_i, \hat{\theta}_{(-i)}, \hat{\delta}_{(-i)})). \tag{2.3.4}$$

Finally, we choose $h = h_*$, where

$$\text{CVAE}(h_*) = \min_{h \in H} \text{CVAE}(h) \tag{2.3.5}$$

and H is a given class of predictor functions. With such a choice of h, we provide the predictor

$$h_*(U, \hat{\theta}, \hat{\delta}) \tag{2.3.6}$$

for future observations, where $\hat{\theta}$, $\hat{\delta}$ is obtained by minimizing the full expression

$$\sum_{j=1}^{n} L(W_j, h_*(U_j, \theta, \delta)). \tag{2.3.7}$$

(The full implication of this procedure is further explained and illustrated with reference to squared error loss in the next section. It may also be mentioned that the estimate $(\hat{\theta}_{(-i)}, \hat{\delta}_{(-i)})$ could have been obtained from the past data omitting the observations on the ith individual in other ways than what is suggested in (2.3.3). Then to determine an optimum predictor, we may have to minimize the CVAE in (2.3.5) with respect to the elements of H as well as the alternative methods of estimating (θ, δ).)

In the subsequent sections, we illustrate the general methods discussed in this section on the chosen data sets with reference to the prediction of a single future measurement Y_{p+1} given the previous measurements $Y_1, \cdots, Y_p$.

Alternative approaches to prediction based on the likelihood principle have been recently developed by Hinkley (1979), Barndorff-Nielsen (1981), Butler (1986) and others. These methods are highly parametric in nature and their applicability in small samples have not been fully examined.

3. LINEAR PREDICTION

In the absence of any information on the stochastic process describing an individual's growth, a standard approach to the prediction problem is to consider the joint distribution of $Y_{p+1} = W$ and $(Y_1, \cdots, Y_p) = U$ over the individuals of the relevant population and derive the conditional distribution of W given U for use in prediction.

In such a case, under quadratic loss function the best predictor is the conditional expectation of W given U, i.e., the regression of W on U, which may involve unknown parameters. The unknown parameters could be estimated from past data and substituted in the regression function to obtain an empirical regression predictor.

When the exact form of the conditional distribution of W given U is not known, it may be possible to restrict the predictor to a given class of functions and estimate it using past data. One such class, which is easy to handle, is the set of linear predictors of the form

$$\hat{Y}_{p+1} = b_0 + b_1 Y_1 + \cdots + b_p Y_p. \tag{3.1}$$

The coefficients $b_0, \cdots, b_p$ are estimated from the past data on n individuals by the least squares method. The value of Y_{p+1} for the current individual with measurements $U_c = (y_{1c}, \cdots, y_{pc})'$ is predicted by

$$\begin{aligned} \hat{W}_c &= \hat{Y}_{p+1,c} \\ &= b_0 + b_1 y_{1c} + \cdots + b_p y_{pc}. \end{aligned} \tag{3.2}$$

The performance of such a predictor has been extensively studied but several problems still remain unsolved. Can we obtain better prediction by considering only a subset of the previous measurements, say by discarding some of the initial measurements? What is the best way of estimating the regression coefficients, by least squares or ridge regression or other methods such as James-Stein's, aimed at reducing the compound mean square error? We shall examine some of these questions.

There are various methods recommended for selection of variables in regression. Excellent reviews can be found in papers by Hocking (1976) and Thompson (1978a, 1978b). In addition we have model selection criteria by Akaike (1973). Most of these are not applicable to the present problem as the sample size is small and the object is not to find a model which explains the observed data but to assess the relative performances of different predictors constructed from past data for use on future individuals. For this purpose the LOO or CV method described in Section 2.3 seems to be more appropriate. This method is outlined below in terms of quadratic loss function.

Let $f(U_c) = f(y_{1c}, \cdots, y_{pc})$ be the true conditional expectation of $Y_{p+1,c}$ given U_c and $\hat{Y}_{p+1,c}$ be as defined in (3.2) through the linear regression of Y_{p+1} on $(Y_1, \cdots, Y_p)$. Then

$$\begin{aligned} &E[(Y_{p+1,c} - \hat{Y}_{p+1,c})^2 \mid U_c, b_0, \cdots, b_p] \\ &\qquad = \sigma^2 + [f(U_c) - \hat{Y}_{p+1,c}]^2, \end{aligned} \tag{3.3}$$

where σ^2 is the conditional variance of $Y_{p+1,c}$ given U_c which may depend on U_c.

Let us consider a subset $(s) = \{Y_{i_1}, \cdots, Y_{i_s}\}$ of the variables $Y_1, \cdots, Y_p$, fit the regression of Y_{p+1} on the subset

$$\hat{Y}_{p+1} = b_0^{(s)} + b_1^{(s)} Y_{i_1} + \cdots + b_s^{(s)} Y_{i_s} \tag{3.4}$$

and estimate $Y_{p+1,c}$ by

$$\hat{Y}^{(s)}_{p+1,c} = b_0^{(s)} + b_1^{(s)} y_{i_1 c} + \cdots b_s^{(s)} y_{i_s c}. \tag{3.5}$$

Then

$$\begin{aligned} &E[(Y_{p+1,c} - \hat{Y}^{(s)}_{p+1,c})^2 \mid U_c, b_0^{(s)}, \cdots, b_s^{(s)}] \\ &\qquad = \sigma^2 + [f(U_c) - \hat{Y}^{(s)}_{p+1,c}]^2. \end{aligned} \tag{3.6}$$

The problem is to choose a subset (s) such that the average of (3.6) over a given distribution of U_c for future individuals is a minimum. One choice of this distribution is the empirical distribution of $U = (Y_1, \cdots, Y_p)'$ based on the past data. In such a case, the average of (3.6) is

$$n^{-1} \sum_1^n \sigma^2(U_i) + n^{-1} \sum_1^n [f(U_i) - \hat{Y}^{(s)}_{p+1,i}]^2. \tag{3.7}$$

In practice, we do not know σ^2 and f and suitable estimates may have to be substituted if we want to use (3.7). An alternative approach is as follows. The CVAE is

$$M(s) = n^{-1} \sum_{j=1}^n (y_{p+1,j} - \hat{Y}^{(s,-j)}_{p+1,j})^2, \tag{3.8}$$

where $y_{p+1,j}$ is the observed value of Y_{p+1} for the jth individual and the second member within the brackets in (3.8) is the predicted value of Y_{p+1} from the regression of Y_{p+1} on the subset $Y_{i_1}, \cdots, Y_{i_s}$ fitted to the past data omitting the observations on the jth individual. The optimum choice of a subset is made by minimizing the CVAE (3.8). Note that in the above procedure no assumption is made about f, the true conditional expectation, and no separate estimate of σ^2 is made. All the other criteria such as Mallows' C_p and the Akaike information criterion described in the references listed above require some assumptions on f and an estimate of σ^2, which introduce additional errors especially in small samples and make the criteria less effective.

The values of $M(s)$ for different subsets of $Y_1, \cdots, Y_p$ are given in Table 5 for the three data sets. It is seen that in all these cases, the regression based on the immediate previous measurement or the immediate two previous measurements provides the best prediction. It appears that the initial measurements do not carry enough information given the later measurements to enhance predictive efficiency by their inclusion in the usual regression analysis.

In the comparisons made in Table 5 all the regressions are estimated by the method of least squares. The LOO method also enables us to examine whether alternative methods of estimation of the regression coefficients such as ridge and James-Stein techniques improve prediction. Computation of the CVAE by using such methods in a few cases showed that the least squares method was more effective for the data sets under consideration.

Note 1. The observed phenomenon that the predictive efficiency is enhanced by omitting the initial measurements can be explained by decomposing the squared multiple correlation coefficient of Y_{p+1} on $Y_p, Y_{p-1}, \cdots, Y_1$ in terms of what are called part correlations (see Rao (1973), page 311),

$$\rho^2_{[p+1][p,\cdots,1]} = \rho^2_{[p+1]p} + \rho^2_{[p+1][(p-1)\cdot p]} + \cdots + \rho^2_{[p+1][1\cdot 2,\cdots,p]} \tag{3.9}$$

where $\rho^2_{[p+1][i\cdot i+1,\cdots,p]}$ is the squared correlation between Y_{p+1} and the residual of Y_i after eliminating the effect of $Y_{i+1}, \cdots, Y_p$. This coefficient, called the part correlation coefficient, is a measure of the improvement in predictive efficiency by the inclusion of Y_i in addition to $Y_{i+1}, \cdots, Y_p$ in the regression analysis. The decompositions (3.9) for the three sets of data were as follows:

Mice data ($p = 6$): $.9082 = .8874 + .0056 + .0019 + .0038 + .0019 + .0076$,

Ramus data ($p = 3$): $.9370 = .9270 + .0100 + .0000$,

Dental data ($p = 3$): $.7448 = .6320 + .1093 + .0035$.

The squared correlation between Y_{p+1} and Y_p dominates in each case indicating that no improvement can be expected by using the other measurements, except perhaps Y_{p-1}. In fact their inclusion introduces more noise and decreases the efficiency of prediction *if* the straightforward regression analysis is used.

Note 2. Having observed that the regression of Y_{p+1} on the previous two measurements Y_p, Y_{p-1} is sufficient for prediction, it may be of interest to examine whether the growth process for an individual can be explained by an *autoregressive type model*

$$Y_{i+1} = \alpha + \beta_1 Y_i + \beta_2 Y_{i-1} + \varepsilon_{i+1}, \qquad i = 2, \cdots, p. \tag{3.10}$$

If the model (3.10) holds, then estimates of α, β_1 and β_2 could be obtained by maximum likelihood or other methods using all the available data. To examine the validity of the model (3.10), separate regressions were fitted in the case of the mice data for Y_7 on Y_6, Y_5; Y_6 on Y_5, Y_4; $\cdots$, Y_3 on Y_2, Y_1. The regression coefficients were as in Table 6 indicating that the model (3.10) may not hold. We could have also

TABLE 5
Cross-validation assessment error of simple linear regression predictor

Previous measurements used	Direct regression	Inverse regression
Mice data (prediction of Y_7, $n = 13$)[a]		
Y_1–Y_6	.095	.103
Y_2–Y_6	.079	.081
Y_3–Y_6	.047	.048
Y_4–Y_6	.037	.040
Y_5–Y_6	.031	.034
Y_6	.027	.028
Ramus data (prediction of Y_4, $n = 20$)		
Y_1–Y_3	.769	.808
Y_2–Y_3	.577	.608
Y_3	.566	.618
Dental data (prediction of Y_4, $n = 27$)		
Y_1–Y_3	4.430	6.211
Y_2–Y_3	3.588	5.227
Y_3	3.665	4.929

[a] The entries are 13 times actual values.

TABLE 6
Regression of Y_{i+1} on Y_i, Y_{i-1}

i	α	β_1	β_2
6	.243	1.045	−.242
5	.035	.806	.261
4	.010	1.280	−.152
3	−.238	1.753	−.107
2	.023	2.192	−1.490

examined the validity of the model (3.10) for predictive purposes by the LOO method.

Note 3 (*inverse regression*). Consider $U = (Y_1, \cdots, Y_p)'$ and $W = Y_{p+1}$ as defined before, and let

$$E(W) = \nu, \qquad E(U) = \mu, \tag{3.11}$$

$$\mathrm{Cov}\begin{pmatrix} W \\ U \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}. \tag{3.12}$$

Then the direct linear regression function which is used for predicting W given U is

$$\hat{W} = \nu + \Sigma_{12}\Sigma_{22}^{-1}(U - \mu). \tag{3.13}$$

Now

$$\begin{aligned} E(\hat{W} \mid W) &= \nu + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\sigma_{11}^{-1}(W - \nu) \\ &= \nu + \rho^2(W - \nu), \end{aligned} \tag{3.14}$$

where ρ^2 is the squared multiple correlation of W on U. Then $W - E(\hat{W} \mid W) = (1 - \rho^2)(W - \nu)$ so that $\hat{W}$ underpredicts W when $W > \nu$ and overpredicts when $W < \nu$. Such a situation may not be desirable in some practical applications. To overcome this feature of the direct regression predictor, an inverse regression predictor

$$\tilde{W} = \nu(1 - \rho^{-2}) + \rho^{-2}\hat{W}, \tag{3.15}$$

which is a modification of the direct regression estimator $\hat{W}$, is sometimes recommended. It is seen that $E(\tilde{W} \mid W) = W$ and $\tilde{W} >, =, < \hat{W}$ according as $W >, =, < \nu$.

The CVAE of prediction obtained by the LOO method using inverse regression is given in Table 5 for various choices of the predictor variables. As expected these values are larger than those for direct regression, but inverse regression provides an insurance against serious bias in prediction.

4. POLYNOMIAL GROWTH CURVES

In Section 3, we have not used any model describing an individual's growth. In the rest of the paper, we shall develop some growth models and study their usefulness in prediction. Let (y_t, t), $t = 1, 2, \cdots$ represent the measurements and the times at which they are taken on an individual. Then models of the type

$$y_t = f(t, \theta) + \varepsilon_t, \quad t = 1, 2, \cdots, \tag{4.1}$$

have been fitted, where ε_t are i.i.d. with a common variance σ^2 and $f(t, \theta)$ is a suitably chosen function of t representing the growth trend depending on an individual specific parameter θ. For a measurement like the stature of an individual, some of the forms of $f(t, \theta)$ recommended are:

$$\begin{aligned} f(t, \theta) &= a + bt - \exp(c + dt) \\ &\qquad \text{[Jenss and Bayley (1937)],} \\ &= k\exp[-\exp(a - bt)] \\ &\qquad \text{[Winsor (1932)],} \\ &= k[1 + \exp(a + bt)]^{-1} \\ &\qquad \text{[Wright (1926)],} \\ &= \sum_{i=1}^{3} k_i[1 + \exp(a_i + b_i t)]^{-1} \\ &\qquad \text{[Bock and Thissen (1976)].} \end{aligned}$$

In fitting models of the above type, the authors were trying to characterize an individual's growth trend over long periods of time. But models which provide an adequate description of the past observations may not necessarily be suitable for predicting future observations. We shall first consider polynomial trends which are simple to fit and which may be adequate to describe growth over a short period of time and examine their usefulness in prediction (see Rao, 1965, 1967, 1976, 1981).

Let

$$y_t = \beta_0 P_0(t) + \cdots + \beta_k P_k(t), \qquad t = 1, 2, \cdots, p, p + 1, \tag{4.2}$$

where $P_0, P_1, \cdots, P_k$ are polynomials over t, which may be chosen to be orthogonal for convenience in computations. Denote

$$X = \begin{pmatrix} P_0(1) & \cdots & P_k(1) \\ \cdot & \cdots & \cdot \\ P_0(p) & \cdots & P_k(p) \end{pmatrix}, \qquad x = (P_0(p + 1), \cdots, P_k(p + 1)), \tag{4.3}$$

$$U_i = (y_{1i}, \cdots, y_{pi}), \qquad W_i = y_{p+1,i}, \qquad \beta_i = (\beta_{0i}, \cdots, \beta_{ki}). \tag{4.4}$$

Then the model for the measurements on the past n individuals can be written in the familiar Gauss-Markoff form

$$\left.\begin{aligned} U_i &= X\beta_i + \varepsilon_i \\ W_i &= x\beta_i + \eta_i \end{aligned}\right\} \quad i = 1, \cdots, n, \tag{4.5}$$

and the incomplete model for the current individual in the form

$$U_c = X\beta_c + \varepsilon_c \tag{4.6}$$

although the variable to be predicted is

$$W_c = x\beta_c + \eta_c. \tag{4.7}$$

The error components in (4.5)–(4.7) are all assumed to be i.i.d. In this section, we shall describe various methods of predicting W_c and compare their relative efficiencies.

4.1 Individual Regression Predictor

In this method we consider the measurements on the current individual only, which are subject to the model

$$U_c = X\beta_c + \varepsilon_c, \tag{4.1.1}$$

and obtain the least squares estimate of β_c,

$$b_c^{(l)} = (X'X)^{-1}X'U_c. \tag{4.1.2}$$

Then, we predict $W_c = x\beta_c + \eta_c$ by

$$\hat{W}_c = xb_c^{(l)}. \tag{4.1.3}$$

The predictor (4.1.3) does not make use of the past complete sets of observations on n individuals explicitly. However, it depends on k, the degree of polynomial fitted and on s the number of immediately previous measurements used. (We use the same notations X, x and U for any selection of the previous measurements.) We shall see later on that it pays to omit (ignore) some of the initial measurements to arrive at a prediction formula. What is the optimum combination of s and k for prediction purposes? We use the past data and apply the CVAE criterion to answer this question.

For the ith individual (in the past data), let $\hat{W}_i(s, k)$ be the predicted value of the $(p + 1)$th measurement using the formulas (4.1.2)–(4.1.3) by fitting a kth degree polynomial to the s previous measurements $Y_{p-s+1,i}, \cdots, Y_{pi}$. Then the CVAE is

$$M(s, k) = n^{-1} \sum_{i=1}^{n} [W_i - \hat{W}_i(s, k)]^2, \tag{4.1.4}$$

where W_i is the observed value of the $(p + 1)$st measurement on the ith individual. We may then choose that combination of (s, k) which minimizes (4.1.4). Table 7, column (3), gives the values of $M(s, k)$ for different values of k and s. It is found that in all the cases studied, the best procedure is to fit a straight line to just the two previous measurements, Y_{p-1}, Y_p, and extrapolate to predict Y_{p+1}.

4.2 Regression on Polynomial Coefficients

Let us consider the past data and compute for each individual the kth degree polynomial regression coefficients fitted to the previous s measurements

$$b_i^{(l)} = (X'X)^{-1}X'U_i, \quad i = 1, \cdots, n, \tag{4.2.1}$$

where X and U_i are the appropriate matrix and vectors for s measurements. Then we have reduced measurements

$$(W_i, b_i^{(l)}), \quad 1, \cdots, n, \tag{4.2.2}$$

computed from the past data, where W_i is a scalar variable and $b_i^{(l)}$ is a $(k + 1)$ vector variable. We fit a multiple regression equation of W on $b^{(l)}$ from the data (4.2.2)

$$\hat{W} = a + g'b^{(l)}, \tag{4.2.3}$$

where a is a scalar and g is a $(k + 1)$-vector and predict W_c by

$$\hat{W}_c = a + g'b_c^{(l)}, \tag{4.2.4}$$

where $b_c^{(l)}$ is obtained in the same way as $b_i^{(l)}$ in (4.2.1) using the same combination of s and k.

The CVAE in such a case is

$$M(s, k) = n^{-1} \sum_{i=1}^{n} (W_i - \hat{W}_i)^2, \tag{4.2.5}$$

where $\hat{W}_i$ is obtained in the same way as $\hat{W}_c$ considering the reduced measurements

$$(W_j, b_j^{(l)}), \quad j = 1, \cdots, i-1, i+1, \cdots, n, \tag{4.2.6}$$

i.e., omitting the combination $(W_i, b_i^{(l)})$. The values of $M(s, k)$ for different values of s and k are given in Table 7, column 4. The values in column 4 are smaller than those in column 3 indicating that the method of regression on polynomial coefficients is better than the method of individual regression prediction.

4.3 Calibration of Individual Predictors

In Section 4.1, we fitted a polynomial of degree k to s previous measurements of an individual and predicted the $(p + 1)$th value by extrapolation. Let $\hat{W}_i$ be the estimate of W_i obtained by the individual regression method for individual i. Then, from the past data we have the pairs

$$(W_i, \hat{W}_i), \quad i = 1, \cdots, n, \tag{4.3.1}$$

from which we can estimate the regression of W on $\hat{W}$,

$$\tilde{W} = a + g\hat{W}, \tag{4.3.2}$$

where a and g are now scalars, and use it to predict W_c,

$$\tilde{W}_c = a + g\hat{W}_c, \tag{4.3.3}$$

where $\hat{W}_c$ is obtained from the measurements on the current individual in the same way as $\hat{W}_i$ for the ith individual.

The CVAE of such a procedure computed by the LOO method using the past data only is given in Table 7, column 5, for different combinations of s and k. There appears to be some improvement over the method of regression on polynomial regression coefficients.

Note that the method employed is similar to that of calibration, which rectifies any deficiency in the

TABLE 7

CVAE of different predictors under the polynomial growth curve model

Previous measurements used (1)	Degree of polynomial fitted (2)	Individual regression predictor (3)	Regression on polynomial coefficients (4)	Calibrated predictor of column 3 (5)	Empirical Bayes predictor (6)
Mice data (prediction of Y_7, $n = 13$)[a]					
Y_1–Y_6	5	7.472	.095	.252	
	4	.600	.076	.235	.375
	3	.175	.058	.093	.139
	2	.104	.060	.037	.087
	1	.206	.049	.035	.194
Y_2–Y_6	4	2.405	.079	.235	
	3	.241	.064	.141	.174
	2	.095	.040	.040	.075
	1	.158	.043	.035	.143
Y_3–Y_6	3	.757	.047	.192	
	2	.096	.039	.052	.069
	1	.111	.039	.034	.097
Y_4–Y_6	2	.229	.037	.094	
	1	.066	.036	.034	.054
Y_5–Y_6	1	.055	.031	.033	
Ramus data (prediction of Y_4, $n = 20$)					
Y_1–Y_3	2	2.989	.769	2.172	
	1	.584	.716	.638	.498
Y_2–Y_3	1	.812	.577	.751	
Dental data (prediction of Y_4, $n = 27$)					
Y_1–Y_3	2	47.398	4.430	9.483	
	1	3.998	3.288	3.680	2.322
Y_2–Y_3	1	12.426	3.585	8.358	

[a] Entries are 13 times the actual values.

directly predicted value due to inadequacy of the assumed model and reduces the overall error.

4.4 Bayes and Empirical Bayes Predictors

The methods developed in Sections 4.1 to 4.3 did not fully exploit the covariance structure of the measurements (U_c, W_c) induced by the linear model (4.6) or (4.7) involving the latent variable β_c. If $E(\beta_c) = \gamma$ and $\operatorname{Cov}(\beta_c) = C(\beta_c) = \Gamma$, then

$$E(U_c) = X\gamma, \qquad E(W_c) = x\gamma, \tag{4.4.1}$$

$$C\begin{pmatrix} U_c \\ W_c \end{pmatrix} = \begin{pmatrix} X\Gamma X' + \sigma^2 I_p & X\Gamma x' \\ x\Gamma X' & x\Gamma x' + \sigma^2 \end{pmatrix} \tag{4.4.2}$$

and the best linear predictor of W_c given U_c is, as derived in Rao (1975),

$$\hat{W}_c = xb_c^{(B)},$$

$$b_c^{(B)} = b_c^{(l)} - \sigma^2(X'X)^{-1} \cdot (\Gamma + \sigma^2(X'X)^{-1})^{-1}(b_c^{(l)} - \gamma), \tag{4.4.3}$$

where $b_c^{(B)}$ is the Bayes estimator of β_c and $b_c^{(l)}$ is the least squares estimator.

The predictor (4.4.3) depends on γ, σ^2 and Γ and also on the choice of the degree of the polynomial and the set of previous measurements used in the analysis. A purely Bayesian approach to the problem when γ, σ^2 and Γ are unknown is considered by Young (1977) and for the choice of the degree of the polynomial by Halpern (1973). There are also Bayesian methods for selection of variables. We do not attempt in this paper to develop a strictly Bayesian approach to deal with all the unknowns. But we shall provide an empirical Bayes approach using the CVAE as the criterion for taking decisions from the available alternatives.

Consider the full model for the ith individual,

$$\begin{pmatrix} U_i \\ W_i \end{pmatrix} = \begin{pmatrix} X \\ x \end{pmatrix}\beta_i + \begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix}, \tag{4.4.4}$$

and obtain the usual least squares estimator and residual sum of squares,

$$b_i^{(l)} \text{ and } S_i^2, \quad i = 1, \cdots, n. \tag{4.4.5}$$

Then we have the following estimating equations when a kth degree polynomial is fitted to the last $(s + 1)$ measurements $Y_{p-s+1}, \cdots, Y_{p+1}$:

$$\hat{\sigma}^2 = [n(s - k)]^{-1}S^2, \qquad S^2 = S_1^2 + \cdots + S_n^2, \tag{4.4.6}$$

$$\hat{\gamma} = n^{-1}(b_1^{(l)} + \cdots + b_n^{(l)}), \tag{4.4.7}$$

$$\hat{\Gamma} = (n - 1)^{-1} \sum_1^n (b_i^{(l)} - \hat{\gamma})(b_i^{(l)} - \hat{\gamma})' - \hat{\sigma}^2(X'X + x'x)^{-1}. \tag{4.4.8}$$

If any diagonal element in $\hat{\Gamma}$ is zero or negative, then the entire row and column defined by the diagonal element are replaced by null vectors. Substituting the estimates (4.4.6) to (4.4.8) in (4.4.3) with some changes in the multiplying constants (see Rao, 1975), we obtain the empirical Bayes predictor

$$\begin{aligned}\hat{W}_c^{(B)} = x[b_c^{(l)} &- g\hat{\sigma}^2(X'X)^{-1} \\ &\cdot (\hat{\Gamma} + \hat{\sigma}^2(X'X)^{-1})^{-1}(b_c^{(l)} - \hat{\gamma})],\end{aligned} \tag{4.4.9}$$

where

$$g = \frac{n(s-k)}{n(s-k)+2}\,\frac{n-k-3}{n-1}.$$

Alternative estimates of σ^2, γ and Γ, which have nice properties, are obtained by ignoring the information on W_i, $i = 1, \cdots, n$. We consider only the incomplete models

$$U_i = X\beta_i + \varepsilon_i, \quad i = 1, \cdots, n, c, \tag{4.4.10}$$

taking U_i as the vector of s measurements y_{ji}, $j = p - s + 1, \cdots, p$ and X as the matrix appropriate to the kth degree polynomial. From (4.4.10), we compute the least squares estimates and the residual sum of squares

$$(b_i^{(l)}, S_i^2), \quad i = 1, \cdots, n, c, \tag{4.4.11}$$

using the same notation as in (4.4.5). Then we have the following unbiased estimating equations:

$$\begin{aligned}\hat{\sigma}^2 &= [(n+1)(s-k-1)]^{-1}S^2, \\ S^2 &= S_1^2 + \cdots + S_n^2 + S_c^2,\end{aligned} \tag{4.4.12}$$

$$\begin{aligned}\hat{\gamma} &= (n+1)^{-1}(b_1^{(l)} + \cdots + b_n^{(l)} + b_c^{(l)}), \\ \hat{\Gamma} &+ \hat{\sigma}(X'X)^{-1}\end{aligned} \tag{4.4.13}$$

$$\begin{aligned}&= n^{-1}B = n^{-1}[(b_c^{(l)} - \hat{\gamma})(b_c^{(l)} - \hat{\gamma})' \\ &\qquad + \sum_{i=1}^{n} (b_i^{(l)} - \hat{\gamma})(b_i^{(l)} - \hat{\gamma})'].\end{aligned} \tag{4.4.14}$$

Substituting these estimates in (4.4.3) with some changes in the multiplying constants (see Rao, 1975), we obtain the empirical Bayes estimator

$$\hat{W}_c^{(B)} = x[b_c^{(l)} - g(X'X)^{-1}S^2B^{-1}(b_c^{(l)} - \hat{\gamma})], \tag{4.4.15}$$

where

$$g = \frac{n-k-3}{(n+1)(s-k-1)+2}. \tag{4.4.16}$$

The CVAE associated with (4.4.15) for different combinations of s and k are given in Table 7, column 6. Although the empirical Bayes predictor shows an improvement over the individual regression predictor, it is somewhat inferior to the method of regression on polynomial regression coefficients and the calibrated predictor.

4.5 Mixed Model

In fitting individual polynomial growth curves, we can allow for the possibility that the coefficients of some of the higher order degree terms are common to all individuals. To deal with such cases in some generality, we consider a mixed model for the ith individual

$$\left.\begin{aligned}U_i &= X_1\xi + X_2\beta_i + \varepsilon_i \\ W_i &= x_1\xi + x_2\beta_i + \eta_i\end{aligned}\right\} \quad i = 1, \cdots, n, c, \tag{4.5.1}$$

where ξ is common to all individuals and β_i varies over the individuals such that $E(\beta_i) = \gamma$ and $C(\beta_i) = \Gamma$. We can assume without loss of generality that

$$\begin{pmatrix} X_1 & X_2 \\ x_1 & x_2 \end{pmatrix} \tag{4.5.2}$$

has orthonormal columns, for otherwise we can make the transformation

$$\xi \to A\xi, \qquad \beta_i \to B\beta_i + C\xi \tag{4.5.3}$$

to ensure the validity of the condition (4.5.2) retaining the non-randomness of ξ. We can then replace the observations on the past individuals by the reduced statistics

$$\begin{aligned}\xi_i^{(l)} &= X_1'U_i + x_1'W_i, \\ b_i^{(l)} &= X_2'U_i + x_2'W_i, \\ S_i^2 &= U_i'U_i + W_i^2 - (\xi_i^{(l)})'(\xi_i^{(l)}) \\ &\qquad - (b_i^{(l)})'(b_i^{(l)}), \qquad i = 1, \cdots, n.\end{aligned} \tag{4.5.4}$$

From these, we obtain the estimates of the unknowns

$$\hat{\xi} = n^{-1}\sum_1^n \xi_i^{(l)}, \qquad \hat{\gamma} = n^{-1}\sum b_i^{(l)}, \tag{4.5.5}$$

$$c_1\hat{\sigma}^2 = \sum_1^n S_i^2 + \sum_1^n (\xi_i^{(l)} - \hat{\xi})'(\xi_i^{(l)} - \hat{\xi}) = S, \tag{4.5.6}$$

$$c_2(\hat{\Gamma} + \hat{\sigma}^2 I) = \sum_1^n (b_i^{(l)} - \hat{\gamma})(b_i^{(l)} - \hat{\gamma})' = B, \tag{4.5.7}$$

where for unbiasedness

$$c_1 = n(p + 1 - a_1 - a_2) + a_1(n-1),$$

$$c_2 = n - 1,$$

a_1 and a_2 being the numbers of the components of ξ and β_i, respectively. From (4.5.6) and (4.5.7) an estimate of Γ is

$$c_2^{-1}B - c_1^{-1}SI, \tag{4.5.8}$$

where if any diagonal element is nonpositive, the corresponding entire row and column are replaced by null vectors.

If all the parameters ξ, γ, σ^2 and Γ are known, then the regression of W_c on U_c is

$$\begin{aligned}\hat{W}_c &= x_1\xi + x_2\gamma + x_2\Gamma X_2'(X_2\Gamma X_2' + \sigma^2 I)^{-1} \\ &\quad \cdot (U_c - X_1\xi - X_2\gamma) \\ &= x_1\xi + x_2\gamma \\ &\quad + x_2[I - \sigma^2 U^{-1}(\Gamma + \sigma^2 V^{-1})^{-1}] \\ &\quad \cdot V^{-1}X_2'(U_c - X_1\xi - X_2\gamma),\end{aligned} \tag{4.5.9}$$

where $V = X_2'X_2$. If the parameters are not known, we substitute the estimates obtained in (4.5.5) to (4.5.8) and obtain an empirical predictor of W_c,

$$\begin{aligned}W_c^{(e)} &= x_1\hat{\xi} + x_2\hat{\gamma} \\ &\quad + x_2[I - g\hat{\sigma}^2 V^{-1}(\hat{\Gamma} + \hat{\sigma}^2 V^{-1})^{-1}] \\ &\quad \cdot V^{-1}X_2'(U_c - X_1\hat{\xi} - X_2\hat{\gamma}),\end{aligned} \tag{4.5.10}$$

where

$$g = \frac{c_1}{c_1 + 2}\,\frac{n - k - 2}{n - 1}. \tag{4.5.11}$$

Note 1. There is some information on ξ, γ, σ^2 and Γ in the observed measurements U_c on the current individual that we have not used. Let $\xi_c^{(l)}$, $b_c^{(l)}$ and S_c be the simple least squares estimators of ξ_c, β_c and the residual sum of squares obtained from the linear model

$$U_c = X_1\xi_c + X_2\beta_c + \varepsilon_c. \tag{4.5.12}$$

If $\xi_c^{(l)}$, $b_c^{(l)}$ and S_c^2 are of the same order of magnitude as $\xi_i^{(l)}$, $b_i^{(l)}$ and S_i^2, $i = 1, \cdots, n$, then improved estimators of ξ, γ and σ^2 could be obtained by combining $\hat{\xi}$, $\hat{\gamma}$, $\hat{\sigma}^2$ with $\xi_c^{(l)}$, $b_c^{(l)}$ and S_c^2 using suitable weights. But the computations would be somewhat complicated.

Note 2. An alternative approach is as follows. We consider only the models

$$U_i = X_1\xi + X_2\beta_i + \varepsilon_i, \quad i = 1, \cdots, n, c, \tag{4.5.13}$$

where X_1 and X_2 may be chosen to satisfy the orthogonality condition

$$\begin{pmatrix} X_1' \\ X_2' \end{pmatrix}(X_1 : X_2) = I. \tag{4.5.14}$$

The reduced statistics in such a case are

$$\begin{aligned}&\xi_i^{(l)} = X_1'U_i, \qquad b_i^{(l)} = X_2'U_i, \\ &S_i^2 = U_i'U_i - (\xi_i^{(l)})'(\xi_i^{(l)}) - (b_i^{(l)})'(b_i^{(l)}), \\ &\qquad\qquad i = 1, \cdots, n, c,\end{aligned} \tag{4.4.15}$$

and unbiased estimates of the unknowns are

$$\begin{aligned}&(n+1)\hat{\gamma} = (n+1)^{-1}(b_1^{(l)} + \cdots + b_n^{(l)} + b_c^{(l)}), \\ &(n+1)\hat{\xi} = (\xi_1^{(l)} + \cdots + \xi_n^{(l)} + \xi_c^{(l)}),\end{aligned} \tag{4.5.16}$$

$$\begin{aligned}n(\hat{\Gamma} + \hat{\sigma}^2 I) &= (b_c^{(l)} - \hat{\gamma})(b_c^{(l)} - \hat{\gamma})' \\ &\quad + \sum_{i=1}^{n} (b_i^{(l)} - \hat{\gamma})(b_i^{(l)} - \hat{\gamma})' = B,\end{aligned} \tag{4.5.17}$$

$$\begin{aligned}&[n + (n+1)(p - a_1 - a_2)]\hat{\sigma}^2 \\ &\quad = S_1^2 + \cdots + S_n^2 + S_c^2 \\ &\quad + \sum_{i=1}^{n,c} (\xi_i^{(l)} - \hat{\xi})'(\xi_i^{(l)} - \hat{\xi}) = S^2.\end{aligned} \tag{4.5.18}$$

Then an empirical Bayes predictor of W_c is

$$W_c^{(e)} = x_1\hat{\xi} + x_2[b_c^{(l)} - gS^2B^{-1}(b_c^{(l)} - \hat{\gamma})], \tag{4.5.19}$$

where

$$g = \frac{n + 1 - a_2 - 2}{n + 2 + (n + 1)(p - a_1 - a_2)}.$$

Then we consider the pairs

$$(W_i, W_i^{(e)}), \quad i = 1, \cdots, n, \tag{4.5.20}$$

where $W_i^{(e)}$ is obtained in the same way as W_c by using $b_i^{(l)}$ in the place of $b_c^{(l)}$, compute the regression of W on $W^{(e)}$,

$$W = a + gW^{(e)}, \tag{4.5.21}$$

and predict W_c by

$$\hat{W}_c = a + gW_c^{(e)} \tag{4.5.22}$$

as in the calibration method discussed in Section 4.3.

5. FACTOR ANALYTIC TYPE GROWTH MODEL

We consider a more general type of growth model (see Rao, 1958)

$$\begin{aligned}Y_{ti} &= B_{1i}\psi_1(t) + \cdots + B_{ki}\psi_k(t) + \varepsilon_{ti}, \\ &\qquad t = 1, \cdots, p, p + 1,\end{aligned} \tag{5.1}$$

where $\psi_1(t)$, $\psi_2(t)$, $\cdots$ are suitably chosen orthogonal functions of time, $\beta_{1i}, \cdots, \beta_{ki}$ are regression parameters specific to individual i and ε_{ti} are independent errors. Writing

$$\begin{aligned}X &= \begin{pmatrix} \psi_1(1) & \cdots & \psi_k(1) \\ \cdot & \cdots & \cdot \\ \psi_1(p) & \cdots & \psi_k(p) \end{pmatrix}, \\ x &= (\psi_1(p + 1), \cdots, \psi_k(p + 1)),\end{aligned} \tag{5.2}$$

$$U_i = (Y_{1i}, \cdots, Y_{pi}), \qquad W_i = Y_{p+1,i}, \tag{5.3}$$

$$\beta_i = (\beta_{1i}, \cdots, \beta_{ki}), \tag{5.4}$$

we can write the models for the past data on n individuals and the current individual in the form

$$\left.\begin{aligned} U_i &= X\beta_i + \varepsilon_i \\ W_i &= x\beta_i + \eta_i \end{aligned}\right\} \; i = 1, \cdots, n, \tag{5.5}$$

$$U_c = X\beta_c + \varepsilon_c \tag{5.6}$$

and the variable to be predicted in the form

$$W_c = x\beta_c + \eta_c. \tag{5.7}$$

The equations (5.5) to (5.7) are of the same form as (4.4.1) to (4.4.3) except that the elements of X and x are not specified. We discuss methods of estimating X and x in addition to the other unknowns and use them in predicting W_c.

5.1 Generalized Principal Components Regression

First we estimate X and $\beta_1, \cdots, \beta_n, \beta_c$ by minimizing

$$\| (U_1 : \cdots : U_n : U_c) - X(\beta_1 : \cdots : \beta_n : \beta_c) \| \tag{5.1.1}$$

for any unitarily invariant norm. This is a well known problem (see Rao, 1980, 1985). This is done by finding the singular value decomposition of the $p \times (n + 1)$ matrix

$$(U_1 : \cdots : U_n : U_c) = \lambda_1 P_1 Q_1' + \cdots \lambda_s P_s Q_s', \tag{5.1.2}$$

where $P_1, P_2, \cdots$ are orthonormal p vectors and $Q_1, Q_2, \cdots$ are orthonormal $(n + 1)$ vectors and $\lambda_1, \lambda_2, \cdots$ are singular values. The optimum choices of X and β_i's are

$$\hat{X} = (P_1 : \cdots : P_k), \tag{5.1.3}$$

$$(\hat\beta_1 : \cdots : \hat\beta_n : \hat\beta_c) = (\lambda_1 Q_1 : \cdots : \lambda_k Q_k)', \tag{5.1.4}$$

where $k \le s$. Then we determine x to minimize

$$\sum_{i=1}^{n} (W_i - x\hat\beta_i)^2, \tag{5.1.5}$$

which yields

$$\hat{x}' = \Lambda^{-2}(W_1\hat\beta_1 + \cdots + W_n\hat\beta_n), \tag{5.1.6}$$

where Λ is a diagonal matrix with $\lambda_1, \cdots, \lambda_k$ as diagonal elements, and predict W_c by

$$\hat{W}_c = \hat{Y}_{p+1,c} = \hat{x}\hat\beta_c. \tag{5.1.7}$$

Note 1. The method described is a generalization of what is known as the principal components regression analysis. The $\hat\beta_i$ may be recognized as the set of k principal components for the ith individual based on the first p measurements.

A slight variation of the method, which is well known, is not to include the observations on the current individual in the estimation of X. We can then develop a general formula applicable to all future individuals. The method can be described as follows. First, we find the singular value decomposition of the $p \times n$ matrix

$$(U_1 : \cdots : U_n) = \lambda_1 P_1 Q_1' + \cdots + \lambda_s P_s Q_s', \tag{5.1.8}$$

in which case X and β_i, $i = 1, \cdots, n$, are estimated by

$$\hat{X} = (P_1 : \cdots : P_k), \tag{5.1.9}$$

$$(\hat\beta_1 : \cdots : \hat\beta_n) = (\lambda_1 Q_1 : \cdots : \lambda_k Q_k)'. \tag{5.1.10}$$

Second, x is estimated by

$$\hat{x} = \Lambda^{-2}(W_1\hat\beta_1 + \cdots + W_n\hat\beta_n) \tag{5.1.11}$$

and, finally W_c is predicted by

$$\hat{W}_c = \hat{x}\hat\beta_c, \qquad \hat\beta_c = (P_1'U_c, \cdots, P_k'U_c). \tag{5.1.12}$$

Note 2. We can build a more general model of the form

$$\left.\begin{aligned} U_i &= \mu + z_i 1 + X\beta_i + \varepsilon_i \\ W_i &= \nu + z_i + x\beta_i + \eta_i \end{aligned}\right\} \qquad i = 1, \cdots, n, c, \tag{5.1.13}$$

where μ is a p vector, z_i and ν are scalars and $1' = (1, \cdots, 1)$. By putting $z_i = 0$, we have the usual principal components model. By putting $\mu = 0$, $\nu = 0$, we have a model with an unknown additive constant. The method of estimation of all the unknowns follows from a general theorem given in Rao (1980, 1985). The formulas developed for the prediction of W_c from the model (5.5) to (5.7) can then be extended to the more general model (5.1.13).

Note 3. Suppose that we have a number of linear models

$$U_i = X\beta_i + \varepsilon_i, \quad i = 1, \cdots, n, \tag{5.1.14}$$

where in addition to $\beta_1, \cdots, \beta_n$ the design matrix X itself is unknown. The method developed in this section enables us to estimate $X, \beta_1, \cdots, \beta_n$ simultaneously. This admits the possibility of testing hypotheses both on X as well as on β_i.

5.2 Factor Analytic Type Regression (Method 1)

Let us consider the model (5.5) to (5.7) and estimate X, x and β_i, $i = 1, \cdots, n$, by minimizing

$$\left\| \begin{pmatrix} U_1 & \cdots & U_n \\ W_1 & \cdots & W_n \end{pmatrix} - \begin{pmatrix} X \\ x \end{pmatrix} (\beta_1 : \cdots : \beta_n) \right\|. \tag{5.2.1}$$

This is achieved through the singular value decomposition

$$\begin{pmatrix} U_1 & \cdots & U_n \\ W_1 & \cdots & W_n \end{pmatrix} = \lambda_1 P_1 Q_1' + \cdots \lambda_s P_s Q_s', \tag{5.2.2}$$

yielding the optimum X, x and β_i,

$$\begin{pmatrix} \hat{X} \\ \hat{x} \end{pmatrix} = (P_1 : \cdots : P_k), \tag{5.2.3}$$

$$(\hat\beta_1 : \cdots : \hat\beta_n) = (\lambda_1 Q_1 : \cdots : \lambda_k Q_k)'. \tag{5.2.4}$$

Then β_c is estimated by the least squares formula

$$\hat\beta_c = (\hat{X}'\hat{X})^{-1}\hat{X}'Y_c \tag{5.2.5}$$

and W_c is predicted by

$$\hat{W}_c = \hat{Y}_{p+1,c} = \hat{x}\hat\beta_c. \tag{5.2.6}$$

As in Note 2, we can apply the above method to the more general model (5.1.13).

Note. There are two steps in the above method: first, that of estimating X and x using the complete sets of measurements on the past individuals; and second, that of estimating the regression coefficients for the current individual for predicting the future values. We can combine these two and formulate our problem as one of finding W_c, X, x, $\beta_1, \cdots, \beta_n, \beta_c$ which minimize

$$(5.2.7) \qquad \left\| \begin{pmatrix} U_1 \cdots U_n U_c \\ W_1 \cdots W_n W_c \end{pmatrix} - \begin{pmatrix} X \\ x \end{pmatrix} (\beta_1 : \cdots : \beta_n : \beta_c) \right\|,$$

where $U_1, \cdots, U_n$, U_c and $W_1, \cdots, W_n$ are known. The optimum value of W_c found by this method is not the same as that given by (5.2.6). A numerical method for solving the problem (5.2.7) is developed in Rao and Boudreau (1985). Further investigation of this problem is currently in progress.

5.3 Factor Analytic Type Regression (Method 2)

Let us consider the model (5.5) to (5.7) with the additional assumption that β_i, $i = 1, \cdots, n, c$, are i.i.d. with mean γ and variance covariance matrix Γ. In such a case

$$(5.3.1) \qquad E\begin{pmatrix} U_i \\ W_i \end{pmatrix} = \begin{pmatrix} X \\ x \end{pmatrix} \gamma,$$

$$(5.3.2) \qquad C\begin{pmatrix} U_i \\ W_i \end{pmatrix} = \begin{pmatrix} X\Gamma X' + \sigma^2 Ip & X\Gamma x' \\ x\Gamma X' & x\Gamma x' + \sigma^2 \end{pmatrix}, \qquad i = 1, \cdots, n, c.$$

Our strategy is to estimate all the unknowns

$$(5.3.3) \qquad X, x, \gamma, \sigma^2 \text{ and } \Gamma$$

from the observations $(U_i' : W_i)$, $i = 1, \cdots, n$, and predict W_c by the usual regression equation

$$(5.3.4) \qquad \hat{W}_c = x\gamma + x\Gamma X'(X\Gamma X' + \sigma^2 I)^{-1}(U_c - x\gamma)$$

with X, x, γ, σ^2 and Γ replaced by their estimates. The estimation of the parameters (5.3.3) have been considered by Bentler (1983) and Sorbom (1974). Because the exact computations are a little complicated, we suggest a simpler method that may be useful in practical applications.

Let S be the corrected sums of squares and products matrix of order $(p + 1) \times (p + 1)$ calculated from the $(p + 1)$-variate samples (past data)

$$(5.3.5) \qquad \begin{pmatrix} U_1 \\ W_1 \end{pmatrix}, \cdots, \begin{pmatrix} U_n \\ W_n \end{pmatrix}$$

and consider the spectral decomposition

$$(5.3.6) \qquad (n-1)^{-1}S = \lambda_1^2 P_1 P_1' + \cdots + \lambda_{p+1}^2 P_{p+1} P_{p+1}$$

with $\lambda_1^2 \geq \cdots \geq \lambda_{p+1}^2$. Then an estimate of (X, x) is given by

$$(5.3.7) \qquad \begin{pmatrix} \hat{X} \\ \hat{x} \end{pmatrix} = (P_1 : \cdots : P_k), \quad k \leq p + 1$$

and those of σ^2 and Γ by

$$(5.3.8) \qquad \hat{\sigma}^2 = (p + 1 - k)^{-1}(\lambda_{k+1}^2 + \cdots + \lambda_{p+1}^2),$$

$$(5.3.9) \qquad \hat{\Gamma} = \begin{bmatrix} \lambda_1^2 - \hat{\sigma}^2 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ & & & & \lambda_k^2 - \hat{\sigma}^2 \end{bmatrix},$$

where the off-diagonal elements in Γ are all zeros and that of γ by

$$(5.3.10) \qquad \hat{\gamma} = (P_1 : \cdots : P_k)' \begin{pmatrix} \bar{U} \\ \bar{W} \end{pmatrix},$$

where $n\bar{U} = U_1 + \cdots + U_n$ and $n\bar{W} = W_1 + \cdots + W_n$.

The CVAEs for some of the methods discussed in this section, computed through the LOO technique on past data, are reported in Table 8.

6. SUMMARY

We considered the problem of predicting a future measurement on an individual given the previous measurements taken at different time points. In practice one could have a large number of previous measurements and their proper utilization poses an interesting problem. Often the required information for forecasting is contained in the previous few measurements or in a small number of summary statistics of previous measurements. In this paper, appropriate models for growth curves are considered to derive the summary statistics, and analyses based on them are described.

Past records on complete sets of measurements enable us through cross-validation or the leave-one-out technique to choose the appropriate prediction

TABLE 8

CVAEs under the factor analytic type model
$(y_t = b_1\psi_1(t) + \cdots + b_k\psi_k(t), t = 1, \cdots, p + 1)$

Principal components	$k = 1$	$k = 2$	$k = 3$
Mice data ($p = 6$, $n = 13$)[a]			
Regression (5.1.6)	.038	.048	.043
Factor analytic Regression (5.2.6)	.038	.061	.071
Ramus data ($p = 4$, $n = 20$)			
Regression (5.1.6)	1.541	.643	.769
Factor analytic Regression (5.2.6)	1.480	.781	

[a] The entries are 13 times the actual values.

function from a given set of alternatives. The CVAE provides the error rate of prediction for future individuals drawn from the same population from which the past individuals could be considered as a random sample. The CVAEs for the different methods of prediction are reported for the three problems considered in the paper.

The methods discussed in the paper are based on a judicious combination of Bayesian and classical techniques in statistics. Some unknown parameters are taken as random variables and some as fixed. The observed measurements are considered as fixed although variations in them are exploited in comparing the performances of different methods.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second Internat. Symp. Information Theory* (B. N. Petrov and F. Csáki, eds.) 267–281. Akadémia Kiado, Budapest.

BARNDORFF-NIELSEN, O. (1981). Likelihood prediction. *Inst. Naz. di Alta Matematica, Symp. Mathematica* **25** 11–24.

BARTLETT, M. S. (1953). Approximate confidence intervals. *Biometrika* **40** 12–19.

BENTLER, P. M. (1983). Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. *Psychometrika* **48** 493–517.

BOCK, R. D. and THISSEN, D. M. (1976). Fitting multicomponent models for growth in stature. *Proc. Ninth Internat. Biometric Conf. Raleigh* **1** 431–442.

BUTLER, R. W. (1986). Predictive likelihood inference with applications (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 1–38.

COX, D. R. (1975). Predictive intervals and empirical Bayes confidence intervals. In *Perspectives in Probability* (J. Gani, ed.) 47–55. Academic, New York.

ELSTON, R. C. and GRIZZLE, J. E. (1962). Estimation of time-response curves and confidence bands. *Biometrics* **18** 148–159.

GEISSER, S. (1975a). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70** 320–328.

GRIZZLE, J. E. and ALLEN, D. M. (1969). Analysis of growth and dose response curves. *Biometrics* **25** 357–382.

HALPERN, E. F. (1973). Polynomial regression from a Bayesian approach. *J. Amer. Statist. Assoc.* **68** 137–143.

HINKLEY, D. V. (1979). Predictive likelihood. *Ann. Statist.* **7** 718–728.

HOCKING, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32** 1–49.

JENSS, R. M. and BAYLEY, N. (1937). A mathematical method for studying growth in children. *Human Biol.* **9** 556–563.

LACHENBRUCH, P. A. (1975). *Discriminant Analysis*. Hafner, New York.

LEE, J. C. and GEISSER, S. (1972). Growth curve prediction. *Sankhyā Ser. A* **34** 393–412.

LEE, J. C. and GEISSER, S. (1975). Applications of growth curve prediction. *Sankhyā Ser. A* **37** 239–256.

POTTHOFF, 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. (1953). Discriminant function for genetic differentiation and selection. *Sankhyā* **12** 229–246.

RAO, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics* **14** 1–17.

RAO, C. R. (1963). Criteria of estimation in large samples. *Sankhyā Ser. A* **25** 189–206.

RAO, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika* **52** 447–458.

RAO, C. R. (1967). Least squares theory using an estimated dispersion matrix and its application to measurement of signals. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 355–372. Univ. California Press.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

RAO, C. R. (1975). Simultaneous estimation of parameters in different linear models and applications to biometric problems. *Biometrics* **31** 545–554.

RAO, C. R. (1976). Prediction of future observations with special reference to linear models. In *Multivariate Analysis IV* (P. R. Krishnaiah, ed.) 193–208. North-Holland, Amsterdam.

RAO, C. R. (1980). Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In *Multivariate Analysis V* (P. R. Krishnaiah, ed.) 3–22. North-Holland, Amsterdam.

RAO, C. R. (1981). Prediction of future observations in polynomial growth curve models. *Proc. Indian Statist. Inst. Golden Jubilee Internat. Conf. 1981, Calcutta* 512–520.

RAO, C. R. (1985). Tests for dimensionality and interactions of mean vectors under general and reducible covariance structures. *J. Multivariate Anal.* **16** 173–184.

RAO, C. R. and BOUDREAU, R. (1985). Prediction of future observations in factor analytic type growth model. In *Multivariate Analysis VI* (P. R. Krishnaiah, ed.) 449–466. North-Holland, Amsterdam.

SORBOM, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British J. Math. Statist. Psych.* **27** 229–239.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 111–133.

STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39** 44–47.

THOMPSON, M. L. (1978a). Selection of variables in multiple regression: Part 1. A review and evaluation. *Internat. Statist. Rev.* **46** 1–19.

THOMPSON, M. L. (1978b). Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples. *Internat. Statist. Rev.* **46** 129–146.

WILLIAMS, J. S. and IZENMAN, A. J. (1981). A class of linear spectral models and analyses for the study of longitudinal data. Technical Report, Dept. Statistics, Colorado State Univ.

WINSOR, C. P. (1932). The Gompertz curve as a growth curve. *Proc. Natl. Acad. Sci. U.S.A.* **18** 1–8.

WRIGHT, S. (1926). Book review. *J. Amer. Statist. Assoc.* **21** 494.

YOUNG, A. S. (1977). A Bayesian approach to prediction using polynomials. *Biometrika* **64** 309–317.