DOCUMENT RESUME

ED 080 589 TM 003 106

AUTHOR	Jones, Paul K.; Novick, Melvin R.
TITLE	Implementation of a Bayesian System for Prediction in m Groups.
REPORT NO	ACT-TB-6
PUB DATE	Jun 72
NOTE	28p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Bayesian Statistics; *Grade Point A.erage; *Mathematical Models; *Prediction; Statistical Analysis; *Student Ability; Technical Reports
IDENTIFIERS	Career Planning Profile

ABSTRACT

A summary of the technical problems encountered in performing Bayesian m group regression is given. Grade-point averages for students entering a vocational-technical program are predicted using ability assessments from the Career Planning Profile (CPP), a development of The American College Testing Program (ACT). The theory derived by Lindley (see Lindley and Smith, in press); a method developed by Jackson Novick, and Thayer (1971); and the cross-validation performed by Novick, Jackson, Thayer, and Cole (in press) are cited; and the relation to conventional least squares is explored. (Author)

FILMED FROM BEST AVAILABLE COPY

ED 080589

ACT TECHNICAL BULLETIN NO. 6

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLIC'

IMPLEMENTATION OF A BAYESIAN SYSTEM FOR PREDICTION IN m GROUPS

by

Paul K. Jones

and

Melvin R. Novick

TM 003 106

The Research and Development Division
The American College Testing Program
P.O. Box 168, Iowa City, Iowa 52240
June 1972

ERIC

Abstract

A summary of the technical problems encountered in performing Bayesian m group regression is given. Grade-point averages for stuients entering a vocational-technical program are predicted using ability assessments from the Career Planning Profile (CPP), a development of The American College Testing Program (ACT). The theory derived by Lindley (see Lindley and Smith, in press); a method developed by Jackson Novick, and Thayer (1971); and the cross-validation performed by Novick, Jackson, Thayer, and Cole (in press) are cited; and the relation to conventional least squares is explored.

Introduction

Least-squares regression equations are commonly used for the prediction of educational performance. Least-squares estimates, unfortunately, are subject to serious sampling fluctuations, especially when the sample size is small. Under certain circumstances, some meaningful improvements are possible. One such situation occurs when the same predictions must be made in several groups. For this case, Lindley (1970) provided a theory for determining simultaneous Bayesian regression equations so that the collateral information available in the other (m - 1) groups improves the prediction equation for each group. Jackson, Novick, and Thayer (1971) provided a detailed method for obtaining the Bayesian estimates. Novick, Jackson, Thayer, and Cole (in press) showed that Lindley's method led, in one study, to a reduction of mean-squared error of prediction or, alternatively, to a saving in sample size when compared to the least-squares estimates in a cross-validation experiment.

In a companion paper, the authors presented Bayesian regression equations computed for 22 vocational-technical programs. Each program was offered by at least six institutions. For our present discussion, we have chosen to focus on 10 institutions who trained students wishing to become machinists. The Machine Work program serves to illustrate the process of predictor selection and that of Bayesian m group regression.

The predictors considered consisted of the seven ability scales of the Career Planning Profile (CPP Form F).

1. Mechanical Reasoning
2. Nonverbal Reasoning

3. Clerical Skills
4. Numerical Computation
5. Mathematical Usage
6. Space Relations
7. Reading Skills

Explanation of the Model

The model proposed bv the Bayesian method ıs identical at the beginning to the classical linear model:

$$\underset{\sim}{Y}_i = \underset{\sim}{X}_i'\underset{\sim}{\beta}_i + \underset{\sim}{e}_i \qquad i = 1, \ldots, m \qquad (1)$$

where:

$\underset{\sim}{Y}_i$ is the n_i-dimensional column vector of freshman grade-point averages at the i-th college.

$\underset{\sim}{X}_i$ is a $(\ell + 1) \times n_i$ matrix deriving from ℓ ability assessments made on n_i individuals.

$\underset{\sim}{\beta}_i$ is a $(\ell + 1)$-dimensional column vector of weights

$\underset{\sim}{e}_i$ is a n_i-dimensional column vector of error components which are normally distributed with a certain residual variance.

The Bavesian model deviates radically, however, in its assumptions regarding the vector $\underset{\sim}{\beta}_i$. Instead of assuming that each $\underset{\sim}{\beta}_i$ is a fixed (but unknown) point in Euclidean space, we assert a probability distribution from which we assume the $\underset{\sim}{\beta}_i$ to be randomly sampled. This assumption can be justified in a Bayesian context by the De Finetti, Hewitt, Savage theorem when our beliefs about the regressions, apriori, are symmetric (exchangeable); that is, our feelings about any one in no way differ from those about any other. In this particular case, the $\underset{\sim}{\beta}_i$ are assumed to behave as a random sample from a multivariate normal distribution. The model is then complete when we express our beliefs about the parameters in the distribution of the $\underset{\sim}{\beta}_i$. A full discussion of the model and of the assumptions regarding distributions on parameters is contained in Novick, Jackson, Thayer, and Cole (in press).

The Bayesian method refers to the assumptions regarding the parameters as prior information. After collecting data, we perform certain mathematical operations to fuse the data and the prior information. If we wish to assign little weight to the prior information, we can do so, and the choice of an appropriately vague prior will be reflected in the fact that the posterior distribution will be based primarily upon present sample information. The form of the prior will typically be chosen to facilitate the integrations or mathematical operations; the practitioner must, as usual, decide whether or not mathematical convenience provides him with an accurate statement of the relationships among his data.

Lindley (1970) has obtained the posterior Bayes density of the regression weights and the residual variances ϕ_i . The joint modal value of this posterior density can be obtained as the solution of 2m linear equations of the form:

$$\phi_i^{-1}(\underset{\sim}{y}_i'\underset{\sim}{X}_i') - \phi_i^{-1}\underset{\sim}{\beta}_i'(\underset{\sim}{X}_i\underset{\sim}{X}_i')$$

$$- (\nu' + m - 1)(\underset{\sim}{\beta}_i' - \underset{\sim}{\beta}_\star')[\nu'\Sigma + (\underset{\sim}{B} - \underset{\sim}{B}_\star)'(\underset{\sim}{B} - \underset{\sim}{B}_\star)]^{-1} = \underset{\sim}{0}' \quad (2)$$

$$- (n_i + 2) + \phi_i^{-1}(\underset{\sim}{y}_i'\underset{\sim}{y}_i) - 2\phi_i^{-1}\underset{\sim}{\beta}_i'(\underset{\sim}{X}_i\underset{\sim}{y}_i) + \phi_i^{-1}\underset{\sim}{\beta}_i'(X_iX_i')\underset{\sim}{\beta}_i$$

$$- \left(\frac{m + 1}{m}\right)\left[1 - \frac{1}{\phi_i(\Theta^{-1} + \kappa)}\right]\left\{\frac{1}{\log[n(\Theta^{-1} + \kappa)]}\right\} = 0 \quad , \quad (3)$$

for i = 2, ..., m

where

$$\underset{\sim}{B} = \begin{bmatrix} \underset{\sim}{\beta}'_1 \\ \underset{\sim}{\beta}'_2 \\ \vdots \\ \underset{\sim}{\beta}'_i \\ \vdots \\ \underset{\sim}{\beta}'_m \end{bmatrix} = \begin{bmatrix} \alpha_1 & \beta_{11} & \beta_{21} & \cdots & \beta_{h1} & \cdots & \beta_{\ell 1} \\ \alpha_2 & \beta_{12} & \beta_{22} & \cdots & \beta_{h2} & \cdots & \beta_{\ell 2} \\ \vdots & \vdots & \vdots & & & & \\ \alpha_1 & \beta_{1i} & \beta_{2i} & \cdots & \beta_{hi} & \cdots & \beta_{\ell i} \\ \vdots & \vdots & \vdots & & & & \\ \alpha_i & \beta_{1i} & \beta_{2i} & \cdots & \beta_{hi} & \cdots & \beta_{\ell i} \end{bmatrix},$$

and

$$\alpha. = \frac{1}{m} \sum_i \dot{\alpha}_i$$

$$\beta_h. = \frac{1}{m} \sum_i \beta_{hi}$$

$$\underset{\sim}{B}_* = \begin{bmatrix} \alpha. & \beta_1. & \cdots & \beta_h. & \cdots & \beta_\ell. \\ \alpha. & \beta_1. & \cdots & \beta_h. & \cdots & \beta_\ell. \\ \vdots & \vdots & & \vdots & & \vdots \\ \alpha. & \beta_1. & \cdots & \beta_h. & \cdots & \beta_\ell. \end{bmatrix}$$

$\theta = m(\Sigma\phi_i^{-1})^{-1}$--the harmonic mean of the ϕ_i, where ϕ_i is the residual in the i-th college.

$\eta = \left(\prod_{i=1}^{m} \phi_i\right) \frac{1}{m}$ --the geometric mean of the ϕ_i .

κ = a small positive constant to be specified.

v' = degrees of freedom constant.

Σ = population covariance matrix for the β_i .

The first set of matrix equations (2) is solved by inserting the starting values and approximating the actual equation with a linear equation; the new values are reinserted to stabilize the solution temporarily. An analogous operation is performed with (3), although here we are dealing with a single scalar ϕ_i . The procedure described above constitutes a single cycle. Normally, 50 cycles suffice for convergence if it is going to occur (at least to the degree of accuracy that interests us). Possible reasons for nonconvergence will be discussed later. The larger the number of predictors, the greater the number of cycles required. Using three predictors for 30 institutions, which represents the maximum number of institutions for which the program is currently dimensioned, fewer than 50 cycles were actually needed.

It is possible to infer from the form of the equations the direction that certain Bayesian analyses will take. Extremely large diagonal entries in the Σ matrix reflect heterogeneity among institutions; this will result in estimates of $\underset{\sim}{\beta}_i$ that are close to the least-squares estimate

$$\hat{\underset{\sim}{\beta}}_i = (\underset{\sim}{X}_i\underset{\sim}{X}_i')^{-1} \underset{\sim}{X}_i\underset{\sim}{y}_i \qquad i - 1, \ldots, m \quad . \qquad (4)$$

Conversely, small diagonal entries in Σ will cause the estimates for those variates to regress immediately to the average weight across institutions. Moreover, the amount of movement of $\underset{\sim}{\beta}_i$ from the least-squares starting point will depend upon the difference between $\underset{\sim}{\beta}_i$ and $\underset{\sim}{\beta}_*$ with outliers being regressed a greater distance than values close to the average value $\underset{\sim}{\beta}_*$.

Similarly, as sample information accumulates for a given institution (assuming m fixed), the growth of the $\underset{\sim}{X}_i\underset{\sim}{X}_i'$ and $\underset{\sim}{X}_i\underset{\sim}{y}_i$ arrays, relative to the other terms in (2), causes the Bayesian estimate to be close to the least-squares estimate. If we have comparatively little information on any given institution but m is of respectable magnitude, the effect is to regress estimates for small institutions toward the average estimate across institutions. Finally, observe that if sample sizes increase dramatically and if the least-squares estimate $\hat{\underset{\sim}{\beta}}_i$ is substituted for $\underset{\sim}{\beta}_i$, then the residual variance estimator ϕ_i approaches the classical estimate.

Procedure

There are five distinct phases of the Bayesian analysis:

1. Formation of sum of cross-products (SCP) matrix.
2. Selection of variables.
3. Production of cards with SCP matrix in required form and using only the selected variables.
4. BPREP program to compute "ideal point" and variance estimates for parameter distributions.
5. Bayesian m group regression (BAYREG) program.

Phases 1 and 3 are basically data processing operations requiring some attention to the details of the model. Phase 2 involves a resource person's talents. Phase 4 is a preparatory step designed to reduce the computing time required for Phase 5 and to transform the data so that the assumptions of the model are more nearly satisfied.

We will now discuss these operations in greater detail.

1. The SCP matrix will be a symmetric array containing the sample size of the institution, the sums of the predictors and the dependent variable, and the sums of cross-products of the predictors and the dependent variable. If ℓ predictors are involved, the SCP matrix will be of order $(\ell + 2)$. For example; tor two predictors, the order will be four, and the matrix will appear as

$$SCP = \begin{bmatrix} n_i & \Sigma x_{i1j} & \Sigma x_{i2j} & \Sigma y_{ij} \\ \Sigma x_{i1j} & \Sigma x_{i1j}^2 & \Sigma x_{i1j} x_{i2j} & \Sigma x_{i1j} y_{ij} \\ \Sigma x_{i2j} & \Sigma x_{i1j} x_{i2j} & \Sigma x_{i2j}^2 & \Sigma x_{i2j} y_{ij} \\ \Sigma y_{ij} & \Sigma x_{i1j} y_{ij} & \Sigma x_{i2j} y_{ij} & \Sigma y_{ij}^2 \end{bmatrix} \quad (5)$$

where the summation is over the n_i students in the i-th school.

Ten institutions offered vocational training for machinists; thus, 10 SCP matrices were created, one for each institution. These matrices collectively summarize all information required for the statistical part of the Bayesian analyses.

Students of matrix theory will recognize that the SCP matrix can be written in partitioned form as

$$\text{SCP} = \begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{X}_i' & \mathbf{1}'\mathbf{y}_i \\ \mathbf{X}_i\mathbf{1} & \mathbf{X}_i\mathbf{X}_i' & \mathbf{X}_i\mathbf{y}_i \\ \mathbf{1}'\mathbf{y}_i & \mathbf{y}_i'\mathbf{X}_i' & \mathbf{y}_i'\mathbf{y}_i \end{bmatrix} \tag{6}$$

where $\mathbf{1}$ designates a column vector containing n_i elements each being unity.

In the application being discussed, seven predictor variables gave rise to a SCP matrix of order nine. Each SCP matrix (one for each college) was then stored on disk file. One advantage of this procedure is that we quickly summarized the data for all future analyses involving these variables and did not need to look at the individual student records again. Another advantage is that by using disk files, we could utilize interactive computer systems. Twenty-two vocational-technical programs were analyzed in all.

2. We were then ready for the selection of variables. First, the resource person was consulted in advance to avoid unnecessary analyses. For example, we would not ordinarily use Mechanical Reasoning to predict social work grades because that skill has no face validity here. However, we might use Numerical Computation if it happened that this skill was necessary to achieve in core courses taken by social work majors.

As a general rule, we used one, two, or at most three predictor variables. Inclusion of more predictors will result in substantial error variation being entered into the prediction. As a rule of thumb, one can remember that for any desired degree of accuracy, an f-fold increase in the number of variables necessitates an f-fold increase in the number of observations. The disadvantage of using too many predictors can be highlighted by the appearance of negative partial regression weights where theory and common sense suggest that there should be none.

A total of only 163 observations spread over 10 colleges was available for the Machine Work program. We should, therefore, choose at most two variables. Consultation with the resource person experienced in the CPP Assessment suggested that the four most likely candidates for predictors were Mechanical Reasoning, Numerical Computation, Space Relations, and Reading Skills. Table 1 reports the results of four variable combinations arranged in the order that the analyses were performed. The least-squares regression slopes are presented, but the intercepts are excluded as they are irrelevant to the problem of predictor selection. The standard deviations of the predictors are similar, so that it is reasonable to study regression weights in place of partial correlations.

The first analysis used Mechanical Reasoning, Numerical Computation, and Space Relations. The Mechanical Reasoning weights fluctuate greatly from one institution to another, but the average weight of .025 is substantial. Similarly, Numerical Computation contributes effectively. Space Relations, however, does not appear to be an effective predictor; if we were to perform Bayesian regression analysis using this set of variables, the Space Relations weights would be close to zero.

Logically, the next step was to substitute Reading Skills for Space Relations. Comparing the average weights, we observe that Numerical

Table 1

Least-Squares Weights for Different Variable Combinations for Machine Work Students

Analysis 1:

School	N	Mechanical	Numerical	Space Relations
1	12	.088	.035	-.058
2	14	-.037	.077	-.031
3	24	-.012	.044	-.001
4	10	.037	.070	-.019
5	12	.027	.065	-.019
6	17	.075	.026	-.002
7	27	.006	.017	.030
8	19	.072	.017	-.003
9	11	.002	.106	-.035
10	17	-.009	.049	-.002
Average Weight		.025	.051	-.014

Analysis 2:

School	N	Mechanical	Numerical	Reading
1	12	-.018	.021	.074
2	14	-.067	.081	.031
3	24	-.011	.045	-.006
4	10	.030	.055	.002
5	12	.022	.057	-.004
6	17	.071	.026	.004
7	27	.003	.005	.040
8	19	.075	.019	-.011
9	11	-.026	.091	.021
10	17	-.009	.055	-.015
Average Weight		.007	.046	.014

Analysis 3:

School	N	Numerical	Reading
1	12	.011	.065
2	14	.054	.007
3	24	.041	-.010
4	10	.063	.016
5	12	.069	.002
6	17	.036	.027
7	27	.004	.041
8	19	.016	.015
9	11	.089	.010
10	17	.051	-.015
Average Weight		.043	.016

Analysis 4:

School	N	Mechanical	Numerical
1	12	.050	.007
2	14	-.060	.083
3	24	-.013	.043
4	10	.031	.056
5	12	.020	.057
6	17	.073	.026
7	27	.018	.019
8	19	.069	.018
9	11	-.011	.091
10	17	-.009	.048
Average Weight		.017	.045

ERIC
Full Text Provided by ERIC

Computation is again the strongest predictor. However, the presence of a large number of negative weights on Mechanical Reasoning and Reading Skills leads us to reject the use of these two predictors together.

Analyses 3 and 4 attempt to assess whether we should retain Mechanical Reasoning or Reading Skills as the second predictor (in addition to Numerical Computation). Examination of the average weights suggests that Mechanical Reasoning and Numerical Computation should be used as predictors though, with samples of the present size, that judgment cannot be at all firm.

For an institution enrolling a very large number of Machine Work students, the final Bayesian prediction equation would be close to the least-squares solution. On the other hand, if an institution has very few students, the weights would be close to the generalized weight equation (GWE) which is roughly the average of the least-squares regression weights across colleges. For schools with modest to large numbers of students in this type of program, a compromise solution between these extremes is automatically provided by the Bayesian analysis. The nature of the compromise reflects the amount of information available from that particular institution and from the other institutions.

Three important points should be made with respect to the GWE:

a) Extreme values (outliers) are regressed more toward the GWE than values close to the GWE.

b) Weights for institutions providing small samples will be regressed more toward the GWE than those for large institutions. This reflects the fact that we place more confidence in estimates computed from a large sample.

c) The GWE is provided as part of the final Bayesian m group analysis. This initial calculation of the average regression weights is intended primarily to aid in selection of variables. One benefit is that it is usually possible to avoid the occurrence of negative weights in the final solution.

3. Once selecticn of variables has been completed, the SCP matrices were transferred from disk file to cards. Before transfer, however, rows and columns containing the predictors that were no longer of interest were deleted.

By this stage, the sheer quantity of cards was no longer a major consideration. Moreover, by using card input in further analysis, it was possible to combine programs where appropr'ate, as well as delete institutions for some programs. For example, a few institutions continued to exhibit negative weights and very low correlations even after m group regression had been completed. Ieletion of these schools had only a slight influence on the GWE.

4. The BPREP program performs several vital functions in preparation for the final run. The most important function is the provision of starting points. Names have been coined for these sets of starting points: they are called least-squares and classical Model II estimates. The classical Model II estimates represent an educated guess as to what the least-squares estimates will look like after Bayesian m group regression has been completed. Roughly, they are some simple weighted averages of the individual regressions with the generalized regressions. These are discussed in detail by Jackson (1971).

BPREP also estimates the variability of the parameters--a necessary procedure in Bayesian prediction. If the underlying parameters vary little from one institution to the next, we tend to estimate the same quantity for each institution. Since unrealistically small, positive, or even negative variance estimates can arise, a lower bound can be set to avoid nonsensical results. Some care is needed here. In some applications.

these variances are really zero or almost zero, and taking too large values for the τ_{β}'s can result in one getting a suboptimum solution if the amount of data is small.

Another function of BPREP is to calculate the "ideal point" at which the predictors should be centered. This point arises from theoretical considerations discussed in Novick, Jackson, Thayer, and Cole (in press). For a single predictor scaled to the ideal point, the slope and intercept parameters are statistically independent, and hence, uncorrelated. The purpose here is to obtain a scale of measurement such that the off-diagonal elements of Σ can be presumed to be very nearly zero.

5. The final step in the procedure involves combining the SCP matrix and the output from BPREP to conduct Bayesian m group regression (BAYREG). This computer program is a relatively sophisticated program designed to solve the simultaneous equations described above. Although the Lindley equations are nonlinear, it is possible to approximate the nonlinear terms using initial estimates of the unknowns. As the system begins to converge, the approximates become more exact.

The principle method of solution is the Crout factoriz. ' ı method. This numerical analysis technique is a modification of the Gauss reduction method and involves the reduction of a linear system to a triangular matrix, followed by successive substitution of each unknown to obtain a solution. A full discussion of this method is to be found in Fox (1964).

Other subroutines in the BAYREG program compute geometric and harmonic means, invert matrices, and compute the inner-dot product of two vectors. Output includes printing of the least-squares and classical Model II starting points as well as the Bayesian solutions resulting from the use of these starting points. When the Bayesian solutions agree for both sets of starting points, we have some assurance that the solution does not depend upon the starting point. Some possible situations which produce lack of agreement are discussed in the next section. In the present investigation, the numerical methods produced slopes which agreed in the first five digits (of which the first three digits were judged significant).

Table 2 reports the least squares and Bayesian estimates for the Machine Work program. For each kind of estimate, we list the number enrolled, the intercept, the slopes for predictors 1 and 4, the estimate of the multiple correlation coefficient, and the residual variance estimate. In addition, we give the generalized weight equation for predicting grades in a Machine Work program at an institution on which we have no information. Note that the generalized weights for $\tilde{\beta}_1$ and $\tilde{\beta}_4$ are close to the average least-square values reported for Mechanical Reasoning and Numerical Computation reported in Table 1.

The slopes for Numerical Computation were regressed toward the generalized weight for this predictor to a greater extent than the slopes for Mechanical Reasoning. This was due to the fact that the variance estimate (tau beta) for the former slope parameter was smaller. As a result, the estimates for $\tilde{\beta}_1$ do not always strictly move in the direction of the generalized weight. We emphasize, however, that none of the Bayesian slope estimates are negative or even close to zero. This strongly suggests that there are true differences in the regressions across the groups and further suggests that an overall pooling method would be suboptimal.

The estimates of the multiple R allow an interesting comparison to be made. The least-squares multiple Rs fluctuate a good deal and almost certainly tend to overestimate the true degree of association. The median of these values is .6340.

ERIC
Full Text Provided by ERIC

Table 2

Least-Squares and Bayesian Estimates for the Machine Work Program

Least-Squares Estimates

School	N	$\hat{\beta}_o$	$\hat{\beta}_1$	$\hat{\beta}_4$	$\hat{R}$	$\hat{\phi}$
1	12	.017	.050	.007	.4749	1.0706
2	14	1.994	-.060	.083	.7609	.3728
3	24	.666	-.013	.043	.6175	.3130
4	10	-1.283	.031	.056	.8524	.4256
5	12	-1.677	.020	.057	.5727	.3833
6	17	-2.690	.073	.026	.7362	.1966
7	27	.801	.018	.019	.3323	.4396
8	19	-1.938	.069	.017	.6505	.4784
9	11	-1.951	-.011	.091	.6695	.7053
10	17	.492	-.009	.048	.4676	.8964

Bayesian Estimates

School	N	$\hat{\beta}_o$	$\hat{\beta}_1$	$\hat{\beta}_4$	$\hat{R}$	$\hat{\phi}$
1	12	- .173	.023	.035	.5900	.3855
2	14	.613	.005	.039	.5931	.3822
3	24	.139	.004	.038	.5925	.3793
4	10	- .573	.032	.036	.5964	.3810
5	12	- .703	.021	.038	.5941	.3803
6	17	- .886	.033	.035	.5954	.3785
7	27	- .089	.021	.033	.5906	.3817
8	19	-1.444	.043	.033	.5931	.3823
9	11	- .307	.010	.040	.5920	.3835
10	17	.067	.006	.039	.5891	.3862
GWE	163	- .336	.020	.037		.3820

The Bayesian estimates of R, however, tend to characterize the program rather than the individual school. Bayesian estimates are computed from the formula

$$\tilde{R} = (1 - \frac{\phi}{\phi_o})^{\frac{1}{2}}$$

where ϕ = residual variance estimate obtained from m group regression with ℓ predictors,

ϕ_o = residual variance estimate obtained by using no predictors.

Finally, the Bayesian estimates of the residual variance tend to cluster. Apparently, the true residual variances are nearly equal for these 10 institutions. The median of the Bayesian estimates is .5928.

Special Technical Problems

1. One important issue is the problem of selection of elements of Σ, the covariance of the intercept and slope parameters for a given program. These parameters are assumed to have a multivariate normal distribution with mean μ and covariance matrix Σ . If the parameters enjoy independent distributions, then Σ will be diagonal.

If the predictors have been scaled to their ideal points, the intercept parameter will be independent of the slope parameters. The ideal point will, in general, be close to the mean of the predictor values. In lieu of further information, it is convenient to assume that Σ is diagonal. Jackson, Novick, and Thayer (1971) demonstrate that moderate departures from this assumption are unimportant.

Once the decision has been made to make Σ a diagonal matrix, the question arises of what to place in the diagonal entries. The Model II estimates of the variances of the slopes and of the intercept from BPREP are given by the following formulas developed by Jackson, Novick, and Thayer (1971).

$$\hat{\tau}_\beta = (m - 1)^{-1}\Sigma(\hat{\beta}_i - \hat{\beta}.)^2 - m^{-1}\Sigma[\hat{\phi}_i/S^2(x_i)] \qquad (7)$$

$$\hat{\hat{\tau}}_\alpha = \frac{\Sigma(\hat{\hat{\alpha}}_i - \hat{\hat{\alpha}}.)^2}{m - 1} \qquad (8)$$

where

$$\hat{\phi}_i = (n_i - 2)^{-1}\sum_j(y_{ij} - \hat{\alpha}_i - \hat{\beta}_i x_{ij})^2 \quad ,$$

$$S^2(x_i) = \sum_j(x_{ij} - x_{i.})^2$$

$$\hat{\hat{\alpha}}_i = y_{i.} - \hat{\hat{\beta}}_i x_{i.} \quad ,$$

$$\hat{\hat{\beta}}_i = \frac{\hat{\beta}_i(1 - m^{-1})\hat{\tau}_\beta + \hat{\beta}.[\hat{\phi}_i/S^2(x_i)]}{(1 - m^{-1})\hat{\tau}_\beta + [\hat{\phi}_i/S^2(x_i)]} \quad .$$

(The dot notation indicates an averaging over the respective quantities.) We refer to these quantities subsequently as tau beta and tau alpha. Note that what is being done here is not strictly Bayesian, but experience suggests that reasonable results will be forthcoming provided one does not ignore real prior information.

It is well known that classical Model II estimates of variance components can be negative. Bayesian m group regression required that a realistic positive estimate be furnished in every case. Since it is unbelievable that the variances are negative or zero, it is appropriate to find some value below which we assign little prior probability for the tau betas.

The choice of a default value (or lower bound) for tau beta was made in the following fashion. The investigators believed from preliminary analyses that the domain of the parameter in question was from 0.0 to 0.05. Assume temporarily that the parameter happens to be distributed uniformly over this interval. The variance is then given by $\frac{c^2}{12}$ where c denotes the interval length. We, therefore, compute

$$\text{approximate } \tau(\beta) = \frac{c^2}{12} = \frac{(.05 - .00)^2}{12}$$

$$= 2(10)^{-4} \quad .$$

As a result of this calculation, we resolve to set 10^{-4} as a lower bound of the variance.

Another investigator may decide that a slightly different default value is appropriate. The important point, however, is that we prevent estimates of tau beta from deviating by more than an order of magnitude from some reasonable norm. There is not much difference between 10^{-4} and $\frac{1}{2}(10)^{-4}$, but there is a great deal of difference between 10^{-4} and 10^{-7} . Again, it would be folly to impose a high apriori value for τ_{β} when the basic situation suggests that this value may be near zero. The particular method used here will be satisfactorv only if we are reasonably certain, apriori, that differences actually exist.

Another difficulty derives from choosing too small a value for tau beta. In this case, the estimates will totally regress to the average value of the starting points; this abnormality conforms to the (possibly) erroneous but possibly correct) assumption that the institutions are identical with respect to this particular regression weight. Moreover, other regression weights will be affected in a multipredictor problem if one or more predictors are totally regressed. Thus, we should avoid setting prior estimates of different tau betas that differ by more than an order of magnitude from each other.

2. Several reasons exist for lack of convergence in the Bayesian m group regression analysis. One possibility is that the posterior distribution is bimodal. In this case, of course, we do not obtain a uniaue modal estimate by differentiating the posterior density and setting the derivative qual to zero. Bimodality may occur in cases in which the sample information is not sufficiently conclusive to overwhelm the mode furnished by the prior density.

In order to certify convergence, the Lindley examinations are routinely solved using both the least-squares and classical Model II regression weights as starting values. This gives some protection against bimodality resulting from the existence of outliers.

The least-squares starting point is, incidentally, to be preferred in the unlikely event that we were predicting grades for m identical colleges, each obeying the model

$$\underset{\sim}{Y} = \underset{\sim}{X}'\underset{\sim}{\beta} + \underset{\sim}{e} \qquad i = 1, \ldots, m \quad . \qquad (9)$$

Examination of the Lindley equations reveals that use of the least-squares starting points result in attaining the unweighted average of the least-squares estimates at the end of the first cycle. If the diagonal elements of Σ are small (i.e., less than 10^{-6}), the system will stabilize at that point for each of the m institutions. Interestingly enough, this beta vector is the least-squares estimator obtained by pooling the information for all m institutions.

Experience has shown that indiscriminate running of more cycles is not a panacea for lack of convergence. In the initial stages of analysis, an investigator may wish to run only 10 cycles; it may be that this is sufficient to guarantee convergence. We do not do this routinely simply because the added cost for computation of the additional cycles is small compared with the bother and cost of rerunning even a small percentage of the analvses. Printing the logarithm of the posterior density function at the end of each cycle is a good means of gauging progress in convergence. Ordinarily, the logarithm will increase dramatically for several iterations until it reaches its classical Model II level.

3. It may be necessary on occasion to delete an institution from the m group regression analysis. Such exclusion should be done only for clearly defined reasons. In this set of analyses, several situations required a decision: First, the schools in the sample had been, in effect,

preselected by their cooperativeness in agreeing to be tested and following established test procedures. Second, the institutions were required to have a certain minimum number of students. For most programs, we stipulated 10 students as the minimum number. In certain programs, however, we had surplus colleges; thus, we increased the required figure to 15 or 20. Third, did the results from the Bayesian m group regression analysis indicate that the college was an outlier? If the partial regression weights were seriously negative for a large school, we concluded that that school was seriously atypical and excluded it. Fortunately this circumstance did not occur frequently.

4. A resource person who is knowledgeable in the substantive area being investigated is in a position to improve the prediction system markedly. For example, suppose that the preliminary statistical analysis causes us to be indifferent in choosing between Numerical Computation and Mathematical Usage. The high correlation between the two tests inclines us to be skeptical about choosing both predictors, and so we are forced to choose between them. The resource person may indicate that one predictor or the other is more appropriate in this case; moreover, a predictor with higher face validity may perform better in cross-validation on a subsequent year's data.

In addition, the advice of the resource person may reduce the number of preliminary analysis required. Once we have established an on-going prediction program, it is not desirable to reconsider all possible subsets of a set of predictors; selectivity is not only more economical but also indicates presence of rational thought.

Running a Bayesian m Group Analysis

The data decks for BPREP and for BAYREG are quite similar, since the data deck for BPREP is a subset of the deck for BAYREG. Accordingly, the data deck for BAYREG is described below:

1. Identification Card		(10A8)
Col. 1-80		Identification for data
2. Parameter Card		(3I4, 3F5.0, 3I2)
Col. 1-4	M	Number of schools (maximum of 30)
5-8	NV	Number of predictors (maximum of 9)
9-12	NCY	Number of cycles (usually 50-200)
13-17	CKAPPA	Small constant (.001)
18-22	DCON	$\lambda = \text{DCON}\ \hat{\tau}$ (normally DCON = 3)
23-27	PHIMIN	Minimum $\hat{\phi}_i$ allowed (can be arbitrarily small)
28-29	IBRAN	0 - $\tau_{\beta\kappa}$'s from BPREP used (typical) 1 - All $\tau_{\beta\kappa}$'s set to some prespecified value 2 - Some $\tau_{\beta\kappa}$'s set to .001
30-31	ISP	1 - Use least square and classical Model II starting values (typical) 2 - Use Bayesian starting values (read in)
32-33	IFN	0 - Log of height of function not printed 1 - Print log of height (typical)
3. Predictor Card		
Col. 1-8		(10A8)
9-16		Name of 1st predictor
.		" " 2nd "
.		.
.		.
4. Scaling Card for Original Scaling Points		(10F8.0)
Col. 1-8		Value to which predictor 1 has been scaled
9-16		Value to which predictor 2 has been scaled
.		.
.		.
.		.
		Value to which criterion has been scaled

5. Scaling Card for Ideal Points (10F8.0)

Col. 1-8	Ideal scaling point for predictor 1
9-16	" " " " predictor 2
.	.
.	.
.	.
	Value criterion has been scaled to

6. Format Card for SCP Matrix (10A8)
The cross products must be read in floating point form.

7. SCP Matrix Cards

For each group, there must be an upper triangular cross-product matrix punched according to the format specified by card 6. The cross-product matrices have the following form for the case of two predictors:

$$\begin{array}{l} \text{Row 1} \\ \\ \text{Row 2} \\ \\ \text{Row 3} \\ \\ \text{Row 4} \end{array} \left[\begin{array}{llll} n_i & \Sigma x_{i1j} & \Sigma x_{i2j} & \Sigma y_{ij} \\ \\ \Sigma x^2_{i1j} & \Sigma x_{i1j}x_{i2j} & \Sigma x_{i1j}y_{ij} & \\ \\ \Sigma x^2_{i2j} & \Sigma x_{i2j}y_{ij} & & \\ \\ \Sigma y^2_{ij} & & & \end{array}\right].$$

Comparison of this matrix with formula (5) reveals that rows 2, 3, and 4 have been translated to the left. These cross products are scaled to the values given by card 4.

8. Coefficient card (5D16.8)

For each group, there must be a coefficient vector and a vector for the squared error of beta. The coefficient vector consists of $(\hat{\phi}_i, \hat{\alpha}_i, \hat{\beta}_{1i}, \ldots, \hat{\beta}^{\sim}_{\ell i})$. The $\hat{\alpha}_i$ is for the data coded to approximate means given by item 3. The squared error of beta vector starts on a new card; it consists of $[S^2(\hat{\beta}_{1i}), \ldots, S^2(\hat{\beta}_{\ell i})]$.

9. Population Variance Estimates Card (5D16.8)

The estimates for population variances are the components of $(\hat{\tau}_{\alpha}, \hat{\tau}_{\beta 1}, \ldots, \hat{\tau}_{\beta \ell}$. If IBRAN (card 2, columns 28-29) is 0, then $\hat{\tau}_{\alpha}$ and $\hat{\tau}_{\beta h}$ are values computed by BPREP program; if some of the original estimates were negative, then these estimates will be set equal to 10^{-3} . If IBRAN equals 1, then all $\hat{\tau}_{\beta h}$ are set to some prespecified value.

The data deck for BPREP differs from that for BAYREG in the following ways. The scaling card for ideal points is not included, nor are the coefficient cards or the population statistics (all these cards result from the BPREP analysis). Summarizing: The respective items in the data decks are listed below.

	BPREP	BAYREG
1. Identification Card	Yes	Yes
2. Parameter Card	Yes	Yes
3. Predictor Card	Yes	Yes
4. Scaling Card for Original Scaling Points	Yes	Yes
5. Scaling Card for Ideal Points	No	Yes
6. Format Card for SCP Matrix	Yes	Yes
7. SCP Matrix Cards	Yes	Yes
8. Coefficient Cards	No	Yes
9. Population Variance Estimates Card	No	Yes

Items 8 and 9 are provided as part of the punched output from BPREP. Item 5, however, must be punched by the investigator and inserted in the data deck (the necessary information regarding the ideal scaling points is contained in the printed output from BPREP).

ERIC
Full Text Provided by ERIC

Several bits of information required in items 1-9 can be placed in the context of Bayesian methodology. The constant CKAPPA bounds the posterior density away from the point at which the residual variances are equal. The value DCON is a factor which inflates the estimates of tau beta by multiplication; thus, it helps ensure that these estimates will not be too small. PHIMIN places a lower bound on the residual variance estimate for any given group.

Often it is desirable to set the value of IFN to one, since the investigator will then be able to observe the computer's search for the mode of the posterior distribution. (The reason that the log of the height function is printed rather than the actual height is that storage of the actual height might cause problems within the computer when the height is very small.) As the iteration process converges, the log of the height function will increase to the modal value and then exhibit very small fluctuations about this value.

The scaling values for the original scaling points (item 4) may be all zero (indicating lack of scaling) if desired; however, if the investigator uses variables with large means, he may wish to scale the variables to their approximate means before producing the SCP matrix. To do this, he simply subtracts a constant from each individual's score. Such a procedure will tend to reduce the magnitude of the numbers contained in the SCP matrix and may decrease the chances for rounding errors in a problem involving large sample sizes. A different scaling point, of course, may be selected for each variable.

The original scaling points should be distinguished from the ideal scaling points. The ideal scaling points are not a mere matter of convenience; they are intrinsic to the Bayesian method being discussed. Also, the ideal points apply only to the predictor variables, whereas, the original scaling points are relevant to the predictors and to the criterion.

We punch the symmetric SCP matrix in a special upper-triangular form. The first row contains the full number of $(\ell + 2)$ elements. If necessary, the punching continues on subsequent cards according to the format card. The second row must begin on a new card but omits the first element in order to begin with the diagonal element in the second row; thus, the punched form of the second row has only $(\ell + 1)$ elements. Continuing in this fashion, we find that the last row, when punched, contains only the diagonal (last) element.

References

Fox, L. An Introduction to Numerical Linear Algebra, Clarendon Press: Oxford, 1964, 117-120.

Jackson, P. H., Novick, M. R., & Thayer, D. T. Estimating regressions in m groups. The British Journal of Mathematical and Statistical Psychology, 1971, 24, 129-153.

Lindley, D. V., & Smith, A. F. M. "Bayesian estimates for the linear model". Journal of the Royal Statistical Society, 1972, in press.

Novick, M. R., Jackson, P. H., Thayer, D. T., & Cole, N. S. "Applications of Bayesian methods to the prediction of educational performance." The British Journal of Mathematical and Statistical Psychology, 1972, in press.

Novick, M. R., Jones, P. K, & Cole, N. S. Predictions of performance in career education. Research Report No. 53. Iowa City, Iowa: The American College Testing Program, 1972.