DOCUMENT RESUME

ED 395 003 TM 024 999

AUTHOR Braun, Henry I.
TITLE Empirical Bayes Methods: A Tool for Exploratory Analysis. Program Statistics Research, Technical Report No. 88-82.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-88-25
PUB DATE May 88
NOTE 61p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Bayesian Statistics; *College Students; Data Analysis; Higher Education; Models; Test Results; *Validity
IDENTIFIERS *Empirical Bayes Estimation; *Exploratory Data Analysis; Graduate Record Examinations; Hierarchical Linear Modeling; Survival Analysis

ABSTRACT

Empirical Bayes (EB) methods are frequently used on hierarchical linear models in practice. This paper provides an overview of parametric EB methods with special emphasis on their application in data-analytic settings. Eight different models with different levels of complexity are described. Comparisons of performance with other methods are illustrated using test data. These data were collected through the Validity Study Service of the Graduate Record Examinations Board for 1980 through 1983. They comprise the records of over 2,000 native English-speaking students in 99 different departments. Applications to validity generalization and survival analysis are also discussed. (Contains 4 tables, 7 figures, and 51 references.) (Author/SLD)

RR-88-25

Empirical Bayes Methods: A Tool for Exploratory Analysis

Henry I. Braun

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 88-82

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

EMPIRICAL BAYES METHODS: A TOOL FOR EXPLORATORY ANALYSIS

Henry I. Braun

Program Statistics Research
Technical Report No. 88-82

Research Report No. 88-25

Educational Testing Service
Princeton, New Jersey 08541-0001

May 1988

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

TABLE OF CONTENTS

Abstract

Empirical Bayes (EB) on hierarchical linear models are frequently utilized in practice. This paper provides an overview of parametric EB methods with special emphasis on their application in data-analytic settings. A variety of models with different levels of complexity are described. Comparisons of performance with other methods are illustrated using test data. Applications to validity generalization and survival analysis are also discussed.

1. Introduction

The models and techniques that fall under the rubric of empirical Bayes (EB) methods constitute an important resource for the analysis and understanding of hierarchical data structures. The goal of this paper is to describe the logic and implementation of a class of EB methods, called parametric EB, and show how they can be viewed as a tool for exploratory data analysis, in a general sense. While most of the illustrations are drawn from the area of educational testing, these methods can and have been employed in a wide variety of settings, some of which are sketched here as well. Nonetheless, this chapter is no eant to be a comprehensive review of all the different applications of EB ideas in recent years.

Why study EB methods at all? The work described here and in the references clearly indicate that EB estimates tend to be more stable and perform better in cross-validation than do classical estimates. A striking instance is given in Braun and Szatrowski (1984) in which it is shown that EB estimates of a set of regression planes are essentially unaffected by differential restriction of range. Other examples are given by Dempster et. al. (1981) and DerSimonian and Laird (1983).

We begin with examples of EB models and a discussion of the considerations underlying the modelling process. Section 3 then presents illustrative EB analyses of data on the prediction of graduate school grades together with typical model comparisons. The next section describes a number of other applications of EB methods to educational measurement, emphasizing the utility of residual analyses. Section 5 touches on validity generalization while Section 6 focusses on survival analysis. Section 7 then deals with a number of miscellaneous issues related to the implementation and interpretation of EB analyses. The final Section 8 provides a brief review of the development of EB methods.

2. Preliminaries

2.1 Setting up an Empirical Bayes Model

We will focus on the simplest two-level structures, where the first level units are nested within second level units. Our prototypical example involves students nested within graduate departments. Suppose then that the data base includes m departments with n_i students in department i. Associated with each student there are measurements of k characteristics, one of which we distinguish as the criterion. In this example, the criterion is first-year average (FYA) and the other characteristics are pre-graduate school measures of academic ability or achievement such as test scores and undergraduate grades. Interest centers on estimating for each department the regression of FYA on the other characteristics. The regression model for department i may be written as

$$Y = X\beta_i + \varepsilon \qquad i = 1,2,\ldots, m \qquad (1)$$

where

Y is an n_ix1 vector of responses,

X is an $n_i$xk matrix of characteristics (including a column of constants)

β_i is a kx1 vector of regression coefficients,

ε is an n_ix1 vector of deviations.

It is usual to assume that

$$\varepsilon \sim N(0,\Sigma),$$

where Σ is a variance-covariance matrix of diagonal form with a common value along the diagonal, usually denoted σ^2. Equation (1) and the associated assumptions are the standard regression setup for which the least squares (LS) estimate $\hat{\beta}_i = (X'X)^{-1}X'Y$ of β_i has many optimality properties.

The situation in practice, however, is somewhat different. Even when the assumption of normality seems appropriate, LS estimates often behave poorly. For example, if n_i is typically small and a longitudinal series of data are available from department i, the $\hat{\beta}_i$ will tend to fluctuate wildly from year to year. The magnitude of the fluctuations does not accord with local expert opinion on changes in the nature of the relation between criterion and predictors and, in fact, these least squares estimates do not perform well in cross-validation. In the educational context, the problem seems to be that various selection processes combine to yield a configuration of data in the predictor space that leads to poorly-determined least squares estimates.

A natural recourse is (to borrow a term from John Tukey) to look for ways to "borrow strength." Can the data from the other departments provide some help in estimating the regression in a given department? One way of formalizing this notion is to assume that

$$\beta_i \sim N(\beta^*, \Sigma^*) \quad \text{(independently).} \qquad (2)$$

The statement (2) implies that the true regression coefficients behave as if they were independently generated from some normal distribution with (unknown) parameters β^* and Σ^*. In Bayesian terminology (2) describes a prior distribution for the β_i. In this setting, the different departments constitute multiple realizations from the prior distribution and, consequently, it is possible to estimate the parameters of the prior.

The equations (1) and (2) jointly constitute an EB model for the data. (It is also referred to as a hierarchical or multilevel linear model.) The standard EB method involves obtaining maximum likelihood estimates (MLE) of β^* and Σ^*, and the

posterior distribution of β_i given the data and these MLEs. The usual EB estimate of β_i, denoted $\tilde{\beta}_i$, is taken to be the mean of this posterior distribution.

It should be noted that in this same situation a true Bayesian would add a third level to the model, namely a presentation of fully specified priors for the parameters β^* and Σ^*. The EB estimates $\tilde{\beta}_i$ may be thought of as approximations to the fully Bayesian estimates. (See Section 7.1).

When the least squares estimates exist, then the EB estimates may be expressed as

$$\tilde{\beta}_i = \left[\frac{\hat{v}_i \hat{\beta}_i}{\hat{v}_i + \hat{w}} + \frac{\hat{w}\ \hat{\beta}^*}{\hat{v}_i + \hat{w}} \right] \qquad (3)$$

where v_i and w are the precisions (reciprocal variances) of the LS estimate and the estimate of the prior mean, respectively. Thus, the EB estimate can be thought of as resulting from "shrinking" the LS estimate toward the estimate of the common mean, with the amount of shrinking depending on the relative precisions of the two estimates. For example, suppose that the data from the department with the most extreme LS estimate is such that the estimate is quite poorly determined. Then the corresponding EB estimate will be pulled in considerably towards the "centre" of the scatterplot of the $\hat{\beta}_i$.

It is a useful fact that EB estimates can be obtained even if the corresponding LS estimate is not uniquely defined. An expression analagous to (3) may be derived in that case (Braun, et. al, 1983).

2.2 Exchangeability

The appropriateness of the EB estimation scheme flowing from (1) and (2) depends critically on the validity of the assumption of exchangeability among the β_i. Essentially, this assumption implies that we have no reason, a priori, to distinguish any one department's vector of regression coefficients from among the others in terms of the values of its components; e.g., that its components should be larger or smaller than the components of any other school. The assumption of independence in (2) is a strong way of implying exchangeability.

In practice, the modelling of exchangeability depends on the extent of our knowledge about the units of analysis and the kinds of measurements we have available to us. For example, if the sample of schools consists of selective chemistry departments, we might be quite comfortable with the assumption of exchangeability among their vectors of regression coefficients. On the other hand, we might well feel uncomfortable with this assumption if the sample of departments were extremely heterogeneous including many different disciplines and different levels of selectivity. One alternative would be to cluster the departments into more homogeneous subgroups and to make the exchangeability assumption separately for each cluster. How to choose the clusters constitutes an interesting problem in exploratory data analysis! Another alternative is to model the departure from exchangeability. That is, if we have some reason to suspect that the size of the regression coefficients for the department depends in some way on measured departmental characteristics, we can try to incorporate this into our model.

Let Z_i be a vector of department-level characteristics for school i. Components of Z_i may include such quantities as the mean test score for students in the department, the size of the class or an indicator for public/private status of the university. We may then write

$$\beta_i = Z_i'\gamma + \delta_i \tag{4}$$

where

$$\delta_i \sim N(0,\Sigma). \tag{5}$$

The assumptions (1), (4) and (5) constitute a new EB model. It postulates that the vectors of true regression coefficients are themselves generated from a regression plane characterized by the matrix γ and independent normal deviations δ governed by the variance-covariance matrix Σ. Note that (5) implies exchangeability among the δ_i; i.e., that all the systematic variation between the β_i has been captured by the regression in (4). Since model (2) is a special case of (4) and (5), the latter can be used to test the adequacy of the simpler model either through formal methods such as the likelihood ratio test for nested hypotheses or through data-based methods such as cross-validation.

Although this model appears slightly more complex, the estimation process is nearly unchanged. MLEs for γ and Σ can be easily obtained and the mean of the posterior distribution for β_i, given the data and these MLEs is taken to be the EB estimate of β_i. The richness of the EB family should now be apparent. Different sets of predictors at the different levels of the model may be tried in various combinations. All the problems encountered in the familiar step-wise regression schemes appear here redoubled, overlaid by the potential for developing clusters of schools for alternative analyses. As we shall see later, the clustering can itself be accommodated in the EB framework.

The task of sorting through all the different models can be a daunting one. One approach to reducing the number of models to be considered is to look at the correlation matrix among potential departmental covariates, discarding those that are contributing redundant information. Another is to run step-wise (multivariate) regressions of the set of LS estimates of β_i on the departmental covariates, eliminating those covariates that do not appear useful. Actually, these regressions are a crude version of the estimation process that a full EB analysis requires. They are faster and should be quite suitable for screening purposes, although more refined procedures are certainly needed here. The EB analysis can then be run on the one or two most promising combinations of covariates. In general, deciding the appropriate level of exchangeability depends on a combination of cross-validation and significance testing.

2.3 Illustrating Empirical Bayes Estimation

Before going on to discuss some analyses of real data, it should prove instructive to examine a schematic which illustrates the consequences for estimation of the different models we have been discussing. Suppose for convenience that we are considering regression through the origin so that β_i consists of a single component and that we have available to us a single school-level covariate that we denote by Z_i. Figure 1 (from Braun and Jones, 1985) displays for eleven departments three estimates of β_i, each plotted against Z_i: the LS estimate and two EB estimates, one derived under the assumptions (1) and (2), the other derived under the assumptions (1), (4) and (5).

In the first case, the EB estimates are equivalent to pulling the LS estimates toward (an estimate of) the point β^* in (2); in the second case, the EB estimates are equivalent to pulling the LS estimates toward the appropriate point on (an estimate of) the line denoted by $Z'\gamma$ in (4). In this illustration there is an apparently strong regression of β_i on Z_i, as suggested by the plot of the LS estimates $\hat{\beta}_i$ against Z_i. Accordingly, for departments with extreme values of Z_i, the two EB estimates result from pulling the LS estimate in different directions. Not surprisingly, then, the exact structure of the EB model can have a substantial effect on the final estimates.

3. An Application of Empirical Bayes

3.1 Data and Models

To illustrate the application of EB methods, we will briefly describe the analysis of some data reported in a slightly different form in Swinton (1986) and Braun, et. al. (1986b). These data were collected through the Validity Study Service (VSS) sponsored by the Graduate Record Examinations Board during the years 1980 through 1983. They comprise the records of over 2000 native English-speaking students at some 99 different departments. Since departments self-select for participation in the VSS, the sample at hand in no way represents a random sample of the universe of graduate departments. Moreover, only departments with ten or more students were included in the study.

The model takes the form:

$$Y_{ij} = \beta_{oi} + \beta_{1i}V_{ij} + \beta_{2i}Q_{ij} + \beta_{3i}U_{ij} + \varepsilon_{ij} \qquad (6)$$

where i indexes graduate departments and j indexes students within departments. V and Q represent scores on the verbal and quantitative sections of the Graduate Record Examination (GRE), rescaled by dividing by 200. Thus the regression coefficients for these variables should be of comparable magnitude to that for undergraduate grade-point average (UGPA), denoted by U in (6), which is on a 0-4 scale. It is usually advisable, for reasons of numerical stability, to rescale the predictors to achieve this comparability.

The criterion, Y, is the first-year average (FYA) in graduate school. It has also been rescaled to be in the range 0-4 for all departments. (It appears to be generally less advantageous to standardize the criterion to have zero mean and unit variance in each department.) The deviations ε_{ij} are assumed to be normally distributed with mean zero and variance σ_i^2. Interest centers on the estimation of the vector of parameters $\beta_i = (\beta_{oi}, \beta_{1i}, \beta_{2i}, \beta_{3i})'$.

For the second level of the model, we assume that

$$\beta_i = Z_i'\gamma + \delta \quad , \qquad (7)$$

where Z_i is a vector of departmental characteristics; namely, the constant and the departmental averages of the three individual-level predictors V,Q and U (denoted $V_{i\cdot}$, $Q_{i\cdot}$, $U_{i\cdot}$). The vector δ is assumed to have a multivariate normal distribution with mean zero and variance-covariance matrix Σ.

Together, (6) and (7) comprise an EB model that matches the hierarchical nature of the data. The conception underlying this model is that the regression coefficients in the prediction equation for a department will depend in some systematic way upon the academic achievements of the department's students as indicated by the three aggregate measures $V_{i\cdot}$, $Q_{i\cdot}$, and $U_{i\cdot}$. Of course, in subsequent explorations various candidate variables may be added to, or deleted from, either of the equations (6) and (7). It should be noted that there is nothing in the logic underlying the model that requires the departmental covariates to match the individual-level predictors as is the case here. In particular, one could exclude one or more of these matched aggregate level covariates and/or include covariates such as department size that have no counterpart on the individual level.

Estimates of the parameters of interest are usually obtained by means of the EM algorithm (Dempster, Laird, Rubin, 1977). The application of EM to EB models is quite straightforward and will not be given here as it is described in a number of sources including the reference above, Dempster et.al. (1981), Braun, et.al. (1983) and Braun and Jones (1985). Some discussion of EM appears in Section 7.2 below.

3.2 Interpreting the Results

Of more immediate concern is how to interpret the results of the estimation process. Table 1 presents an estimate of the matrix γ, based on (6) and (7). One interpretation of $\hat{\gamma}$ is that a typical department with mean test scores $\bar{v}, \bar{q}$, and $\bar{u}$ among its matriculants should have regression coefficients for the constant, V, Q and U given, approximately by:

$$\beta_0{}^* = 1.53 + 1.39\bar{v} - .50\bar{q} - .49\bar{u}$$

$$\beta_1{}^* = .86 + .11\bar{v} - .09\bar{q} - .25\bar{u}$$

$$\beta_2{}^* = .31 - .21\bar{v} + .04\bar{q} + .07\bar{u}$$

$$\beta_3{}^* = -.18 - .33\bar{v} + .10\bar{q} + .31\bar{u}$$

For a department with precisely these mean test scores, $\hat{\beta}_i$, the LS estimate of β_i, is pulled toward $\hat{\beta}^*$ to obtain the EB estimate, $\tilde{\beta}_i$. For example, one department in our sample recorded $\bar{v} = 2.49$, $\bar{q} = 2.49$, and $\bar{u} = 3.17$. We find that

$$\beta^* = \begin{bmatrix} 2.19 \\ .12 \\ .11 \\ .23 \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} .96 \\ -.10 \\ .46 \\ .44 \end{bmatrix} \qquad \tilde{\beta} = \begin{bmatrix} 1.33 \\ .09 \\ .21 \\ .38 \end{bmatrix}$$

Note that each component of $\tilde{\beta}$ lies between the corresponding components of β^* and $\hat{\beta}$. This need not always occur, however. In this case, the negative LS coefficient for verbal is slightly positive in the EB estimate. This often happens in the estimation of prediction equations for graduate departments as is seen in Figures 2 and 3, where we present scatterplots for the 99 departments of LS estimates and UGPA, respectively. In the latter case, while there are many negative LS estimates, all the EB estimates are positive. In the former case, some EB estimates are negative, but much less so than the corresponding LS estimates. These figures also illustrate rather dramatically the reduced variability among the EB estimates in comparison to the variability among the LS estimates. It is interesting to note that the estimated variance components in Σ^* are rather small: for β_1 it is 3.7×10^{-4} and for β_3 it is 1.6×10^{-2}. The data seem to suggest then, that there is little variability about the plane.

It is difficult to gauge from inspection of the $\hat{\gamma}$ matrix how strong is the apparent relationship between covariates and regression coefficients. One approach is simply to compute β^* for different combinations of covariates and to see how much they differ. For example, another department recorded $\bar{v} = 3.09$, $\bar{q} = 3.15$ and $\bar{u} = 3.46$. For this department $\beta^{*\prime} = (2.55 \quad .05 \quad .03 \quad .19)$. Its prediction plane is more elevated, but shallower, than the one presented just before. Another approach is to determine whether any plausible combinations of covariate values lead to a β^* with negative components.

3.3 Alternative Models

The fact that we can obtain numerical estimates of the parameters of the model is no guarantee that we can not do better. The first step in exploring the space of models is to experiment both with different sets of predictors and with different sets of covariates, imposing different structural assumptions on the data. For example, we might want to add such covariates as the variances of the test scores among matriculants in the department since the magnitude of the regression coefficients in the various departmental prediction equations may well be affected by differential restriction of range. Below we will display some comparisons among competing models of this sort.

We can also adopt another strategy of model criticism. In the context of our example, we have made a rather extraordinary assumption; namely, that for our purposes the enormous heterogeneity of graduate disciplines and departments can adequately captured by a few simple aggregate measures of student preparation. To put it another way, under the model an economics department and a physics department with comparable students, as measured by average test scores and UGPA, would be expected to have similar prediction equations. The labels economics and physics are considered to contain no useful information. Actually, this runs counter to current practice in which data is pooled over all departments in a given discipline, or even over a number of related disciplines, and a single prediction equation estimated by least squares.

We can explore some alternatives in this direction by clustering disciplines according to various subjective criteria and then fitting an empirical Bayes model of the form (6) and (7) separately to the departmental data from each cluster. Thus the assumption here would be that while departmental labels within clusters are not informative, the cluster labels are. To illustrate, we may divide the graduate disciplines into five clusters: Humanities, Social Sciences, Psychology, Biological Sciences, Physical Sciences and Engineering (see Braun and Jones, 1985 for more details). Empirical Bayes models can then be fit to each cluster and the results compared to models involving no clustering. To add interest to the competition, we may add another cluster-based model in which the prediction equations for all departments in the same cluster are constrained to have the same slopes, but intercepts are allowed to vary arbitrarily. The set of equations for each cluster are then fit by least squares.

3.4 Cross-Validation of Models

How are the comparisons to be carried out? Since the purpose of the estimation process is to develop an instrument for prediction, it seems most appropriate to employ cross-validation (Stone, 1978). Ordinarily, the sample is divided in half with the model estimated on one-half (the calibration sample) and the predictions validated on the other half (the validation sample). In this case, with many of the departments so small, it seems more sensible to set aside a small fraction of the sample for validation leaving most of the data from each department for model estimation. For this exercise, three students in each department were set aside and the model estimated on the basis of the remaining $\Sigma(n_i-3)$ observations.

Results of a cross-validation exercise can be reported in many ways. In the areas of measurement and testing, the correlation of observed with predicted is a favorite summary statistic. Here, however, we prefer to focus on the residuals themselves; that is, for each department we use the estimate of its prediction equation to predict the FYAs of the three students set aside for validation and compare these predictions to the FYAs actually observed. Following standard statistical practice, we define the residual to be observed minus predicted.

For illustrative purposes, we compare the performance of eight different models. Except for the first, each model yields an equation of the form (6) for each department. These models are described below:

OM: The mean FYA in the calibration sample from the department is used as the predictor.

LSD: Ordinary LS estimate, using data from only that department.

LSC: LS estimates generated by discipline cluster; within a cluster, departments have common slopes but different intercepts.

LSA: LS estimates treating all 99 departments as a single cluster; departments have common slopes but different intercepts.

EB1: EB using departmental predictor means as covariates in (7); single analysis incorporating all disciplines.

EB1C: As EB1, but model fit separately to each discipline cluster.

EB2: As EB1 but including variances of predictors as additional covariates in (7).

EB2C: As EB1C, but including variances of predictors as additional covariates in (7).

The LSC method corresponds to carrying out an analysis of covariance (ANACOVA) separately in each cluster. The EB methods represent a generalization of the standard ANACOVA since they allow different slopes as well as different intercepts. The various EB approaches simply postulate different models for the variability among departmental slopes.

Table 2 presents a summary of the performance of these six models, using the mean squared error of prediction (average of the squared residuals), denoted MSE, as the criterion. The first column presents the results aggregated over all 99 departments, or 3x99 = 297 predictions. Except for OM and LS, all the methods perform quite similarly, with EB1 having a slight edge. The remaining columns present the results separately for each of the five discipline - clusters.

There are several points worth noting. First, it is somewhat surprising that using the overall mean is superior to using LS. This is eloquent testimony to the volatility of the latter procedure, apparent especially in the results for the Biology cluster. Second, EB1 and EB2, which do not use cluster information, generally outperform EB1C and EB2C, which do - even when the results are displayed by cluster. Thus, this particular choice of clusters does not seem to aid estimation. On the other hand, LSC does rely on the clusters and performs quite well. This suggests that the size of the departments in each cluster needs to be somewhat larger before we can reliably distinguish differences in slopes. Finally, we note that EB2 does no better than EB1 even though it employs additional covariates that might plausibly be related to the magnitude of the regression coefficients in a department. Thus the most parsimonious EB model is to be preferred. In fact, additional evidence suggests that a single covariate should usually suffice.

A more sobering view of this exercise is to compute the square root of the typical MSE, the root-mean-square-deviation, which for EB1 is approximiately 0.35. Thus, the RMSD is approximately one-fourth to one-fifth the typical range of FYAs in a department. It is somewhat disheartening that all our efforts can not reduce uncertainty in prediction to a greater extent. Of course, we have not explored other choices of predictors and covariates that might yield some improvements.

3.5 More on Clustering

It should be noted that the fitting of EB models separately to different clusters can be brought fully within the EB framework by explicitly recognizing this third level of the hierarchy. Specifically, (3) implies for the ith department in the kth cluster that

$$\beta_{ik} = Z'_{ik}\gamma_k + \delta_{ik} \quad \delta_{ik} \sim N(0, \Sigma). \tag{8}$$

We then add the assumption that

$$\gamma_k' \sim N(0, \xi \otimes \tau) \tag{9}$$

where ξ and τ are matrices and $\otimes$ denotes the Kornecker product. As far as I know, such a model has not been implemented in practice, at least not in the EB framework. Another alternative is to experiment with forming different sets of clusters of departments and fitting EB models separately to each of the new clusters. This was carried out in Braun and Jones (1985), using the distribution of GRE subject test scores as the basis for clustering departments. The resulting prediction equations, based on five empirically determined clusters, did not offer any improvement either over the global EB model (no clustering). or the discipline-based clusters already described. In other settings, however, alternative clusters could lead to improved estimates.

4. Other Applications of Empirical Bayes

4.1 Introduction

In the previous sections we have seen how the empirical Bayes paradigm provides a rich family of models with which to model hierarchical data. There is, perhaps, an embarrassment of riches since in many instances it can be extremely time-consuming to study even a fraction of the plausible models. Nonetheless, with the aid of cross-validation and other diagnostics, it is usually possible to select a serviceable model without an inordinate expenditure of effort. In this section we illustrate how empirical Bayes models can be used in a variety of ways to facilitate exploratory analyses.

4.2 Cross-Stratification of the Population

It often happens that the population under study can be classified in different ways. For example, in the education context, students can be classified both by the school they attend and their ethnicity. We may be interested in how both these factors affect the relation between the criterion and the predictors. Such an instance arose in a study of the predictive validity of the Graduate Management Admissions Test (GMAT) for White and Black students (Braun, et.al. 1983).

The aim of this investigation was to explore differential predictive validity. Unfortunately, Black students comprised only four percent of the sample of 8500 drawn from 59 schools. The modal number of Black students at a school was two and only eight schools had ten or more Black students enrolled. Using classical methods it would be clearly infeasible to estimate separate prediction equations for Black and White students at each school. However, the EB methodology does make such a goal practicable. The model for school i takes the form:

$$Y_{ij} = Z_{ij}[\beta_{1i} + I_{ij}\beta_{2i}] + \varepsilon_{ij}\ ,\ \varepsilon_{ij} \sim N(0,\sigma_i^2) \text{ independently,}$$

where

$$Z_{ij} = (1\ V_{ij}\ Q_{ij}\ U_{ij})$$

$$I_{ij} = \begin{cases} 1, & \text{if the student j is Black} \\ 0, & \text{if the student j is White.} \end{cases}$$

Here V_{ij} and Q_{ij} denote the student's scores on the verbal and quantitative sections of the GMAT and U_{ij} denotes the UGPA.

This model does provide for separate regression planes for White and Black students in each school, characterized by the vectors of coefficients β_{1i} and $\beta_{1i} + \beta_{2i}$, respectively. We then assume that $\beta_i = (\beta'_{1i}\beta'_{2i})'$ is governed by the distribution:

$$\beta_i \sim N(\beta^*, \Sigma^*)\ .$$

This setup facilitates the borrowing of information in two directions: across departments within race and across race within departments. The fitted models proved quite stable and informative comparisons among prediction equations were carried out, even when there was insufficient data in the department to obtain LS estimates of the prediction equations. The interested reader is encouraged to read Braun, et.al. (1983) or Braun and Jones (1981) for further details.

When one of the classifications yields a dominant group (in terms of sample size) and several smaller groups, estimating prediction equations separately for each group may not be desirable. In such a case, a simple residual analysis may be sufficient. Braun, et.al. (1986a) studied the question of whether test scores obtained by disabled students taking special administrations of the SAT predicted first year college grades for those students as well as did test scores obtained by non-disabled students taking regular administrations of the SAT. Special administrations may simply involve allowing the student extra time or presenting the examination in a different format (large type, Braille or cassette) or both. Students with disabilities are usually divided into four categories: hearing impaired, visually impaired, learning disabled and physically handicapped. Except for the first group, these students tend not to cluster at specific schools.

For this study, we used EB methods to estimate a set of college-specific prediction equations based on data from regular test administrations. These equations were used to generate residuals both for non-disabled and disabled students. An example is given in Table 3.

It is evident that while the residuals for the non-disabled are relatively well-behaved, those for the disabled students are not, indicating some differential validity. In particular, note that the trend in mean residuals with increasing levels of predicted FYA (rows 5, 6, 7). Subsequent analyses suggested that the anomalous results for the hearing impaired students were due to grading practices in the two special schools many attended while for the learning disabled students they were largely due to effects of allowing excessive additional time. The simple residual analysis was sufficient to lead the investigators into productive lines of inquiry.

4.3 Empirical Bayes Models for Extrapolation

The above analysis, it should be admitted, could probably have been accomplished using least squares estimates since for most schools the sample size of the baseline group was substantial. When the baseline group is not large, the use of EB estimates should confer substantial advantages in yielding informative residuals. In the next example, however, the use of EB methods seems mandatory.

The object of this study (Braun, et.al. 1986b) was to investigate the predictive validity of GRE test scores obtained in special administrations. As one might expect, the test volume is very small and very few students attending graduate school have taken special administrations of the GRE. Those that have are scattered across a variety of departments in hundreds of different schools. The principal obstacle to carrying out a residual analysis similar to the one described for the SAT was that baseline data was generally not available. That is, the vast majority of the departments where the disabled students matriculated had not participated in the VSS offered by the Graduate Record

Examination Board, so that ETS had no data on other matriculated students for those departments. Considerations of time and money precluded embarking on a second massive data collection project following on one that had been required in order to obtain criterion data for the disabled students.

The remedy was to employ a variant of the empirical Bayes models already mentioned to obtain indirect estimates of departmental prediction equations. Briefly, the same set of departments employed in the analysis in Section 2 was used to fit an empirical Bayes model of the form (6) and (7). In this application, however, there were two new covariates, replacing the ones used previously: the means of the GRE-V and GRE-Q among students who had their scores sent to the particular department, rather than among matriculants to the department (since the latter were unavailable). Through cross-validation we were able to show that predictions based on this model behaved as well statistically as those from more conventional models. We could then turn our attention to those departments where the disabled students had matriculated. For those departments, as well, we had data available on score-senders and substituting the score-sender means into the fitted version of equation (7) yielded an estimate of the regression coefficients for the department's prediction equation. In the language of Section 2, our estimate corresponds to shrinking the least squares estimate of the prediction equation (which is unavailable here) all the way to the plane defined by equation (7). The key here was the recognition that the little auxiliary information available for departments could be used to calibrate an EB model that would yield estimates of the desired quantities.

With estimated prediction equations in hand, a residual analysis for the first-year grades of disabled students who had taken a special administration of the GRE was carried out along the lines already described for the SAT. Although the residuals were somewhat noisier than before (see Table 4), the same general patterns emerged, lending some credence to the approach.

5. Validity Generalization

5.1 Introduction

Meta-analysis (Glass, 1976; Light and Pillemer, 1984; Hedges and Olkin, 1985) is a set of techniques that were developed to facilitate combining inferences across different studies of the same, or related, phenomena. A paradigmatic example is a set of studies undertaken in different classes to examine the efficacy of a new program relative to the standard.

In the "fixed effects" approach, a generalized linear model is constructed relating observed treatment effects to various study characteristics, with the aim of investigating the nature of the association between true treatment differences across studies and differences across studies on the included characteristics. These characteristics might be such qualitative factors as the sex of the teacher or the grade level of the class while quantitative factors might be the size of the class or the mean class score on a pre-treatment test. One outcome of such an analysis is adjusted estimates of the true treatment effects derived from the fitted version of the model. (See Rosenthal and Rubin, 1982; Hedges and Olkin, 1983.)

In the "random effects" approach, the emphasis lies in decomposing the observed variance among treatment effects (now treated as realizations from some distribution) into components that can be attributed to different sources of variation. (See Rubin, 1981; DerSimonian and Laird, 1983.) The EB approach corresponds to a "mixed model" setup with both fixed and random effects. Raudenbush and Bryk (1985) provide a clear exposition of this sort of analysis.

In this section, we will focus on a special case of meta-analysis, termed validity generalization (VG). Here interest centers on investigating the variation in validity coefficients among different studies with the aim of ascertaining what proportion of the variation may be attributed to "artifactual" sources such as differences in sample sizes, criterion reliability, restriction of range, etc. In the employment testing context, see, for example, Hunter, et.al. (1982) or Schmidt (1987). The latter presents a summary of the work of one group of investigators who are convinced that nearly always almost all the observed variation is artifactual (This view is not shared universally.) In the educational context, VG is examined by Linn, et.al. (1981) and by Linn and Hastings (1984). Interestingly, these authors conclude that there is a substantial VG in the context of predicting first year law school grades but that there is strong evidence for at least some situational specificity.

Traditionally, VG studies have emphasized the random effects approach. This is due, in part, to the perspective adopted by Schmidt, Hunter and their collaborators: If essentially all the variation among validity coefficients is artifactual, there is no point in building regression models for them. In fact, the overall mean will serve as the best estimate of the true validity in each study. This view represents one end of the continuum spanned by EB models.

5.2 The EB Approach

Hedges (1987) nicely demonstrates how EB methods can be usefully applied to the VG setting, especially when there are missing data. Hedges deals specifically with psychometric aspects of the problem, particularly with the problem of correcting the Z's for unreliability and restriction of range. He demonstrates how when missing data precludes the calculation in all studies of these corrections, the simplicity and power of the EM algorithm show to good advantage.

Consider a set of n studies from which correlation coefficients r_1, r_2, ..., r_n are obtained. Let $\hat{T}_i$ represent a version of r_i ($i = 1, 2, \ldots, n$) corrected for restriction of range and reliability. Then define T_i to be the Fisher Z-transform of $\hat{T}i$:

$$T_i = 1/2 \log [(1 + \hat{T}_i)/(1 - \hat{T}_i)] .$$

We suppose that

$$T_i \sim N(\Theta_i, \sigma_i^2) \qquad (10)$$

where Θ_i is the true transformed validity in study i. Taking a random effects approach, we may further suppose that

$$\Theta_i \sim N(\Theta^*, \Sigma^*) . \qquad (11)$$

Expressions (10) and (11) are a slightly simplified version of the model (1) and (2) with which we introduced EB methods.

The Schmidt-Hunter view corresponds to making the inference $\Sigma^* = 0$ and, consequently, employing $\hat{\Theta}^*$ as an estimate of the common value of all the Θ_i. However, if the data suggest that $\Sigma^* \neq 0$, then separate estimates of Θ_i are called for.

Hedges' focus is on developing improved point estimates of the underlying correlations through the use of EB methods. It should be noted, however, that the empirical distribution function of the EB estimates of validity is not a good estimate of the distribution of the true validities. The latter generally shows more dispersion than the former. This was pointed out by Louis (1984), among others, and he indicates how different EB estimators are required if the principal purpose of the exercise is to estimate the distribution of true validities rather than to develop optimal estimates for the validity in each department. These insights must be pursued in order to develop a true EB analog to the variance components analysis that is now standard in VG.

The remainder of this section will deal with two issues that arise in introducing EB into this area. We will not be specifically concerned with trying to estimate the degree of VG that can be inferred for a particular data set, but rather how different are the statistical estimates arising from different procedures. The first issue is how much of a difference the use of EB can make. We will compare two procedures based on the GRE data already introduced. No corrections for reliability or restriction of range will be applied here.

The first procedure regresses the departmental validity coefficients, $\hat{r}_i$, arising from fitting equation (6), on a set of six departmental covariates comprising the means and variances of the predictors among the students in the department. Let $\hat{w}_i$ denote the fitted value of the validity coefficient for department i resulting from the fitted plane. (This is the basis of the procedure adopted by Linn and Hastings in their approach to VG.)

The second procedure begins by computing

$$\hat{Z}_i = 1/2 \log \; [(1+\hat{r}_i)/(1-\hat{r}_i)]$$

and carries out an EB analysis based on the model:

$$\hat{Z}_i \sim N(\Theta_i, (n_i-3)^{-1})$$

$$\Theta_i = X'\gamma + \delta \; , \; \delta \sim N(0, \Sigma) \; ,$$

where X contains the same departmental covariates included in the first procedure. Let $\tilde{\Theta}_i$ denote the resulting EB estimates of Θ_i and let

$$\tilde{v}_i = \exp\{2\tilde{\Theta}_i - 1\}/\exp\{2\tilde{\Theta} + 1\} .$$

The quantity $\tilde{v}_i$ is just the inverse Fisher transform of the EB estimate of $\tilde{\Theta}_i$. Figure 4 plots $\hat{w}_i$ against $\tilde{v}_i$ and it is evident that the two estimates are very close for most departments although as one might expect, the distribution of the $\hat{w}_i$ is more short-tailed than that of the $\tilde{v}_i$. In this case, EB has not made much of a difference. Of course, as Hedges points out, it can be of crucial importance when data are missing.

The second issue is the possibly different meanings that can be placed on the results of carrying out an EB analysis on different levels. For example, suppose we obtain EB estimates of the regression coefficients β_i in each department using the model (6) and (7), again employing the same departmental covariates as in the two procedures above. Denote these estimates by $\tilde{\beta}_i$ as usual. Now compute

$$\tilde{r}_i^{\,2} = (\tilde{\beta}_i' \, \Sigma_i \, \tilde{\beta}_i)/(\tilde{\beta}_i' \, \Sigma_i \, \tilde{\beta}_i + \tilde{\sigma}_i^{\,2}) \quad , \qquad (12)$$

where Σ_i is the variance-covariance matrix of the predictors among students in department i and $\tilde{\sigma}^2{}_i$ is the EB estimate of the residual variance about the regression plane. Figure 5 plots $\tilde{r}_i$ against $\tilde{v}_i$ and it is evident that $\tilde{r}_i$ tends to be substantially smaller than $\tilde{v}_i$. In fact median $(\tilde{r}_i) = 0.41$, while median $(\tilde{v}_i) = 0.58$. What accounts for this difference?

My own interpretation is that $\tilde{v}_i$ represents an "adjusted" estimate of concurrent validity while $\tilde{r}_i$ represents an "adjusted" estimate of predictive validity. That is, suppose we were able to generate a second set of data for each department, independently of the first, and having the same Σ_i as before. Then $\tilde{v}_i$ is an estimate of the concurrent validity we would observe in that second sample. On the other hand, $\tilde{r}_i$ estimates the correlation we would observe between the criterion data in the second sample and predictions based on using $\tilde{\beta}_i$, derived from the first sample. Thus the drop from 0.58 to 0.41 represents the typical (for this data set) attenuation in validity in moving from a concurrent to a predictive mode. Thus these two quantities are really answering different questions.

Note that replacing $\tilde{\beta}_i$ with the LS estimate of $\hat{\beta}_i$ of β_i in (12) yields the ordinary R^2-statistic. This would generally be a poor predictor of how well $\hat{\beta}_i$ would predict the criterion in the second sample. The derived validities $\tilde{v}_i$ could themselves be subjected to an EB analysis to determine the degree of VG for predictive validity. An interesting question arises if the more complex models corresponding to (1), (4) and (5) are employed. What study-level characteristics are suitable candidates for inclusion as covariates in the higher level of the model? The answer may not be the same as when we use EB models simply to obtain improved estimates of validity. I believe that more attention needs to be paid to the nature of the process that VG is meant to illuminate and that current approaches may be inadequate in this regard.

6. More General Applications of EB

6.1 Miscellaneous Examples

EB ideas have been extended by now to numerous areas of application and to many classical procedures, as a perusal of the Current Index to Statistics quickly indicates. Laird (1978) has shown how to incorporate EB methods in the estimation of models for two-way contingency tables while Mislevy (1987) has applied it to the estimation of item parameters in item response theory models. Mason and Wong (1985) have shown how the EB paradigm can be applied to the case of logistic regression; i.e., when the criterion to be predicted takes the values 0 and 1. They too assume normal priors of the form (7) for the vectors of coefficients resulting from the logistic regressions. Unfortunately, the estimation procedures are somewhat more complicated, principally because there are no sufficient statistics available. Consequently, the EM algorithm requires successive passes through the data, which can become expensive for large data sets. Wong (1986) has indicated some simplifications may be possible.

Another approach, similar to that suggested at the end of Section 2.2, may be useful here. Suppose the model takes the form:

$$\text{logit } P = X'\beta_i \tag{13}$$

$$\beta_i = Z_i'\gamma + \delta_i$$

$$\delta_i \sim N(0, \Sigma^*) \quad \text{independently.}$$

In the demographic example described by Wong and Mason, the event of interest was whether a woman had ever used a modern contraceptive. Individuals were grouped by country (here, i indexes country), and predictors included in X were level of education and type of residence during childhood. The country level covariates were Gross National Product and an index of the effectiveness of the national family planning program.

The suggestion is to obtain first the ordinary logistic regression estimates $\hat{\beta}_i$ of β_i along with the estimated variances $\hat{\sigma}_i^2$ of these estimates.

Equation (1) can then be replaced by:

$$\hat{\beta}_i \sim N(\beta_i, \hat{\sigma}_i^2) \quad \text{, independently.} \qquad (14)$$

EB estimates of β_i can be derived from (14), (4) and (5) using the standard EM algorithm. While this procedure cannot be fully efficient, it should serve as useful screening device. The more burdensome Wong-Mason method then need only be applied to a few selected combinations of predictors and covariates, for which the approximate procedure can provide useful starting points.

6.2 Classical Survival Analysis

One area in which EB methods should, perhaps, play more of a role is survival analysis. Especially in medical research, sample sizes tend to be rather small. Consequently, survival curves and, especially, hazard functions are rather poorly estimated. Most of the theoretical Bayesian work, however, has focussed on the estimation of a single survival curve (see Phadia (1980) for an extensive review). Very little in the Bayesian context has been done on borrowing information across several samples.

In the frequentist domain, however, considerable progress has been made through the use of generalized linear models. Suppose individuals are clustered into I homogeneous groups which can be characterized by a vector of covariates, Z. Assume for convenience that the components of Z are all indicator functions. The first level of the model postulates that events in group i are governed by a hazard function $\lambda_i(\cdot)$. The second level postulates that $\lambda_i(t) = \lambda_o(t) \exp\{Z_i'\beta\}$ where $\lambda_o(\cdot)$ represents a baseline hazard function and β is a vector of coefficients to be estimated. This model was proposed by Cox (1972) in a now-classic paper. Cox's interest centered on the estimation of β and the comparison of different choices for Z. Somewhat surprisingly, he showed that β could be estimated without specifying the form of $\lambda_o(\cdot)$ using a "partial likelihood" approach. Although Cox's justification has been criticized, he has offered an alternative derivation (Cox, 1975). Estimation can be carried out using GLIM (Whitehead, 1980), even if certain parametric forms for $\lambda_o(\cdot)$ are specified (Aitken and Clayton, 1980).

Another approach has been suggested by Holford (1980) and Laird and Olivier (1981). They assume that $\lambda_o(\cdot)$ can be approximated by a piecewise exponential hazard (step-function) over a suitably chosen set of intervals. They then show how the resulting estimation problem is formally similar to one involving the estimation of log-linear models for contingency tables incorporating Poisson data, a problem that can be easily solved using existing software such as LOGLIN or GLIM.

From our point of view, these methods facilitate the borrowing of information across groups resulting in more stable estimates of hazard functions than could be obtained using the data from a single group alone. (Not surprisingly, differences in the estimates at the level of the survival curve are usually not very large - integration is a wonderful smoother!)

6.3 EB Survival Analysis

The Bayesian perspective could be introduced in a number of ways. One would be to combine the log-linear model representation with the work of Laird (1978) already mentioned. A somewhat simpler tack would be to adapt the already established normal theory methods to this problem. This approach has been worked out (Braun 1985) and is sketched briefly here.

Suppose that there are K groups of individuals and that data is collected for T-time intervals of equal length. For each cell in the group x time matrix we require two pieces of information: the total exposure (measured in person-years or equivalent units) and the number of events that occurred. Let

e_{ik} = amount of exposure for individuals in group k during time interval i and

d_{ik} = number of events occurring for individuals in group k during time interval i.

We then assume that conditional on e_{ik}, the distribution of d_{ik} is Poisson with parameter $\Theta_{ik}\, e_{ik}$. The unknown Θ_{ik} represents the (constant) hazard rate assumed to be operating during interval i for group k.

The classic estimator of Θ_{ik} is d_{ik}/e_{ik}, the so-called occurrence-exposure rate. These estimates tend to be quite unstable. In order to bring the usual EB machinery to bear, we transform the problem.

Define

$$X_{ik} = [(d_{ik} + .375)/e_{ik}]^{1/2} .$$

Conditioning on the matrix of exposures, we assume

$$X_k \sim N(\mu_k, S_k) \qquad (15)$$

where

$$X_k = (X_{1k} \ldots\ldots X_{Tk})'$$

$$\mu_k = (\Theta_{1k}^{1/2} \ldots\ldots \Theta_{Tk}^{1/2})'$$

and

S_k is a diagonal matrix with the i^{th} diagonal element being $(4e_{ik})^{-1}$. The second level of the model assumes that the μ_k are independently generated from some multivariate normal distribution; i.e.

$$\mu_k \sim N(\mu, \Sigma), \text{ independently} \qquad (16)$$

If the groups conform to a factorial structure or if they can be characterized by numerical covariates, (16) could be replaced by a model of the form(4) and (5):

$$\mu_k = Z'\gamma + \delta$$

$$\delta \sim N(0,\Sigma), \text{ independently.}$$

Note that the model does not make strong assumptions about the shape of the underlying hazard function. The data are allowed to determine the best approximating step-function, which can then suggest particular parametric forms for subsequent analyses. These models include the one described in Section 6.2 as a special case in which there is no stochastic component at the second level. As presently formulated, our models seem to be related to doubly stochastic Poisson process models (Grandell, 1972).

There are two features of this application, one minor and one major, that distinguish it from others already discussed. The minor one is that S_k, the variance of X_k, is fixed and need not be reestimated during the course of the iterations of the EM algorithm. The major feature is that Σ generally contains too many parameters, particularly when k is small and T is relatively large. Our solution has been to constrain Σ to take the following form:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots \\ \rho & 1 & \rho & \cdots \\ \rho^2 & \rho & 1 & \cdots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{pmatrix}.$$

The assumption of geometrically decreasing correlations seems reasonable for this application. Because formulas for the MLEs of σ^2 and ρ can be expressed in terms of a cubic equation (Szatrowski, 1976), only a few modifications of the basic computational algorithm are required.

To illustrate the effect of EB estimation, we present two figures from Braun (1985) based on a reanalysis of a retrospective survival study (Cutait, Lesser and Enker, 1983) of the effectiveness of prophylactic oophorectomy in patients with cancer of the large bowel. There were 308 patients distributed among four groups according to whether or not they had had an oophorectomy and based on the stage of the disease (Duke staging, levels B or C). The analysis was restricted to durations up to seventy-two months beyond the original surgery and this interval was divided into twelve six-month intervals. There were 110 deaths during the period of the study.

Figure 6 displays the classical estimates of the true hazard rates while Figure 7 displays the EB estimates. The latter are evidently much better behaved although this alone does not establish their superiority. For this, variations on the cross-validation methodology are required and these are described in Braun (1985).

The model proposed here, including the assumption of a patterned structure for Σ, should also be suitable for the analysis of repeated measures designs. In that case X_{ik} would represent the i^{th} measurement on the k^{th} individual.

7. Miscellanea

7.1 Empirical Bayes vs. Bayes

It has already been mentioned that the parametric EB models we have considered here are closely related, in a formal sense at least, to fully Bayesian procedures. In the EB framework, inferences about the parameters of interest are made conditional on the observed data and the MLEs of the parameters of the prior. The latter are obtained either directly through recourse to the marginal distribution of the observables or recursively through the EM algorithm.

In a proper Bayesian analysis, fully specified hyperpriors for these parameters would instead be proposed and a standard Bayesian solution would be developed. It has been argued (Rubin, 1981) that in normal models the EB solution ordinarily represents a convenient approximation to the fully Bayesian approach, provided that the likelihood function for the prior parameters is nearly symmetric about a point in the interior of the parameter space and that non-informative hyperpriors are employed.

In the example Rubin presented, however, the likelihood function for variance parameter of the prior did achieve its maximum on the boundary of its range. As a consequence, the EB inferences substantially underestimated the variability among the true parameter values. Rubin's solution was to carry out a summary Bayesian analysis using Monte Carlo methods.

A similar problem arose in the EB estimation of survival curves (Braun, 1985). The MLE of the parameter ρ (see Section 6.3) took the value unity - on the boundary of its range. A fully Bayesian analysis was carried out but the resulting inferences proved not very different from those derived from the EB analysis. The lesson, perhaps, is that the investigator should always be vigilant for anomalies in the likelihood function that might be indicative of a problem with the EB approach.

It should be recalled that even under the best of circumstances, care must be taken to obtain valid estimates of the uncertainty surrounding EB estimates. This point is addressed very well by Morris (1983).

The connection between Bayes and EB methods are also addressed by Deeley and Lindley (1981). However, they are concerned with the formulation of the EB problem proposed by Robbins (1955) which differs from that discussed here. (See Section 8).

7.2 Robustness

One issue that has received comparatively little attention is that of robustness of EB procedures. In Robbins' formulation the prior distribution is estimated from the data, so that the question of the sensitivity of the inferences to the assumed form of the prior does not arise.

In the parametric EB setup we have emphasized here, such questions do arise. Unfortunately, since it is so convenient to use the conjugate normal prior, very little investigation has been carried out.

Some discussion of robustness appears in Morris (1983) and the accompanying discussion. There is no general agreement except that when the number of units is small, nonparametric estimation of the prior is unlikely to be useful. Leonard (1983) indicates that substantially different estimates can emerge when nonparametric estimation techniques are employed. The key issue in practice is what is lost when the distribution of the unknown parameters is long-tailed but modelled by a symmetric prior. Berger (1983) suggests that EB should be quite robust since misspecification of the prior would ordinarily lead to minimal shrinkage.

Laird (1982) has carried out some preliminary studies of the effectiveness of employing a nonparametric maximum likelihood estimate of the prior, while Laird and Louis (1982) discuss a related problem in the more general context of incomplete data problems.

Interestingly, a rather complete analysis in the Poisson problem has been recently carried out by Gaver and O'Muircheartaigh (1987). They find that the EB estimates of event rates are relatively insensitive to the choice of priors considered. Of course, much more work needs to be done in this area.

Returning to the regression problems that we have presented, EB techniques give evidence of being fairly resistant to outliers. That is, a few aberrant observations at the individual level usually do not have a substantial impact on the EB estimates of the corresponding regression plants. In that case, the effect of borrowing of information overwhelms the information provided by those data.

7.3 EM Algorithm

We have not commented much on numerical considerations in obtaining EB estimates of families of parameters, except to say that the EM algorithm provides a convenient method. While it is easy to implement in this setting, convergence of EM can be slow even in relatively small problems and the computation of estimated variances is not automatic. (This is perhaps less serious in the EB context.) In typical problems at ETS we often run 500 iterations to assure convergence.

A number of authors have suggested improving the speed of EM by incorporating some features of a Newton-Raphson algorithm in the process. Louis (1982) and Meilijson (1986) have developed such procedures but they have not been applied to EB problems. See also Laird et. al. (1987).

8. Brief Review of the Development of EB

The name Empirical Bayes has been attached to two related but different statistical methods. The term was coined by Robbins (1955) who introduced it in the context of developing optimal sequential decision rules within a Bayesian framework. Robbins' interest was in procedures which did not require explicit estimation of the underlying prior distribution. A excellent review of early work was given by Maritz (1970) while more recent work by Cressie (1982) helps to characterize those problems that are amenable to Robbins' approach. A second school was established by Efron and Morris (1973, 1975) who developed the insights gained from Stein-estimators (James and Stein, 1961) into a set of techniques for the simultaneous estimation of many parameters. The connection between empirical Bayes (EB) and Bayesian techniques made more explicit in Deeley & Lindley (1981) and in Morris (1983). The latter provides a review and informative discussion. The latter also treats the problem of interval estimation in the EB context and provides references to a number of interesting applications of EB methodology to real-world problems.

Rubin (1980, 1981) has emphasized the notion of EB solutions as convenient approximations to fully Bayesian analyses and the importance of checking the reasonableness of the approximation through examination of the appropriate likelihood function. The popularity enjoyed by the EM algorithm in EB calculations is due largely to Rubin's influence.

Dempster, Rubin and Tsutakawa (1981) discussed the application of EB to more complicated covariance component models while Braun et.al. (1983) treated the problem of estimating regressions when the population of units can be classified according to a factorial structure in which many cells are sparsely populated or even empty. Braun and Jones (1985) explicated EB models for vectors of regression coefficients that incorporated regression models in the prior. Similar models for univariate study effects were presented by Raudenbusch and Bryk (1985).

1.53	1.39	-.50	-.49
.86	.11	-.09	-.25
.31	-.21	.04	.07
-.18	-.33	.10	.31

Table 1: Estimate of γ in (6)

	All (99)	Humanities (12)	Social Science (43)	Psychology (10)	Biological Science (16)	Physical Science (18)
OM	.15	.18	.14	.14	.15	.16
LSD	.19	.18	.19	.14	.30	.16
LSC	.13	.16	.13	.11	.14	.14
LSA	.13	.16	.13	.11	.13	.14
EB1	.12	.16	.12	.10	.11	.13
EB1C	.13	.17	.13	.11	.12	.16
EB2	.12.	.16	.12	.10	.10	.14
EB2C	.14	.15	.13	.15	.13	.17

Table 2: Cross-validation Estimates of Mean Squared Error of Prediction for Eight Models. Number of departments in parentheses.

Row		Nonhandi-capped Controls	Disabilities							
			Hearing		Learning		Physical		Visual	
			Standard	Special	Standard	Special	Standard	Special	Standard	Special
1	Number	6255	130	84	99	437	198	72	35	171
	Means									
2	Actual FYA	0.00	-0.06	-0.24	-0.38	-0.49	0.00	-0.19	0.20	-0.11
3	Predicted FYA	0.00	-0.31	-0.51	-0.41	-0.42	-0.04	-0.08	0.06	-0.16
4	Residual	0.00	0.25	0.27	0.03	-0.07	0.04	-0.11	0.14	0.05
	Residuals									
5	Low Predicted	.03	.52	.74	.12	.14	.15	.21	.28	.31
6	Med. Predicted	-.07	.02	.25	.04	-.03	-.03	-.26	-.20	.02
7	High Predicted	.04	.21	-.20	-.07	-.31	.02	-.23	.27	-.18
	Standard Deviations									
8	Actual FYA	1.00	1.08	0.96	1.12	1.00	0.95	1.07	1.00	1.06
9	Predicted FYA	0.50	0.56	0.60	0.50	0.50	0.54	0.52	0.40	0.55
10	Residual	0.37	1.00	1.01	1.07	0.96	0.82	1.01	0.93	1.00
	Correlations									
11	Actual & Pred.	.49	.39	.23	.33	.34	.50	.35	.37	.37

Table 3: Residual Analysis for Disabled College Students

First year average predicted by SATs and HSGPA. Standard refers to disabled students taking regular administrations of the SAT. Special refers to disabled students taking special administrations of the SAT. Rows 5, 6 and 7 present mean residuals conditioned on whether the predicted FYA fell in 1st, 2nd or 3rd tercile of distribution of FYAS.

	Nonhandicapped	Handicapped				
		Standard		Special		
			Total	Learning	Physical	Visual
1. Number	2025	184	216	19	48	105
Means						
2. Actual FYA	3.48	3.40	3.38	3.49	3.46	3.31
3. Predicted FYA	3.50	3.46	3.47	3.42	3.50	3.47
4. Residual	-0.02	-0.06	-0.09	0.07	-0.04	-0.16
Mean Residuals						
5. Low Predicted	-0.06	0.10	-0.04	0.10	-0.08	0.06
6. Medium Predicted	-0.00	-0.10	-0.02	-0.31	0.12	-0.11
7. High Predicted	0.02	-0.15	-0.20	-0.22	-0.16	-0.28
Standard Deviations						
8. Actual FYA	0.42	0.50	0.52	0.48	0.55	0.54
9. Predicted FYA	0.23	0.20	0.20	0.20	0.16	0.20
10. Residuals	0.33	0.49	0.51	0.53	0.49	0.53
Correlations						
11. Actual & Predicted	0.63	0.24	0.27	0.23	-0.04	0.29

Table 4: Residual Analysis for Graduate School Disabled Students.

First Year Average (FYA) predicted by GREs and UGPA. Standard refers to disabled students taking regular administrations of the GRE. Special refers to disabled students taking special administrations of the GRE. Rows 5, 6 and 7 present mean residuals conditioned on whether the predicted FYA fell in 1st, 2nd or 3rd tercile of distribution of FYAS.

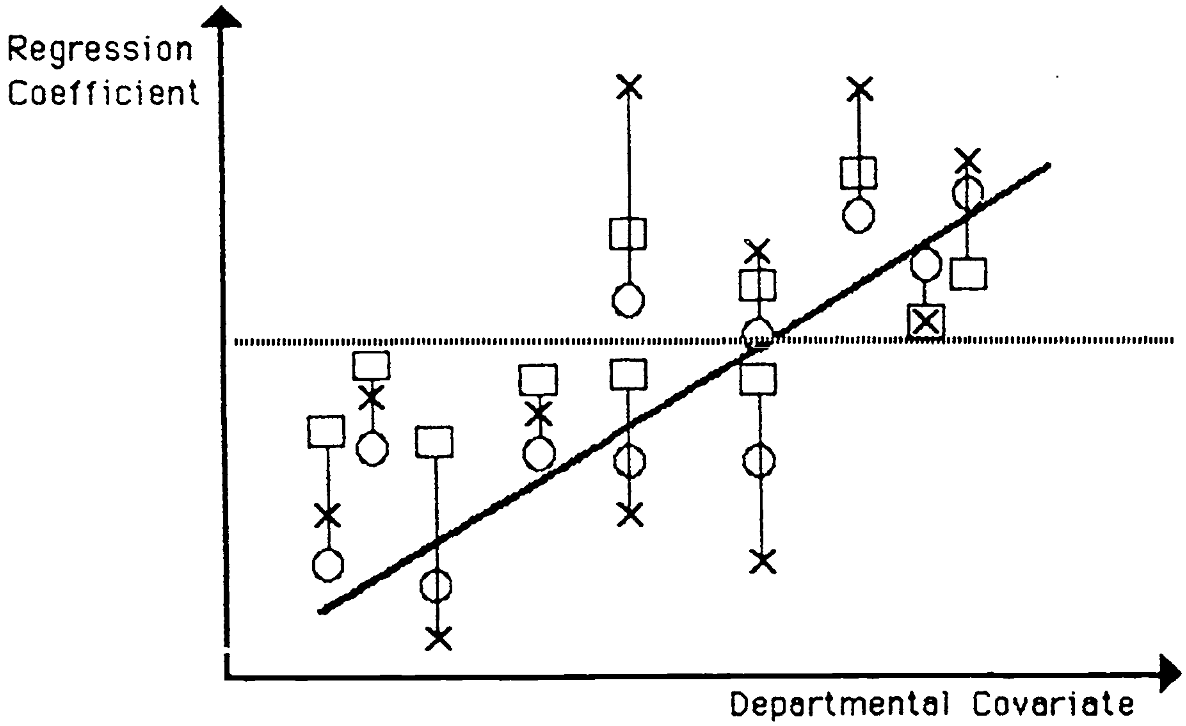

Figure 1: Effects of Empirical Bayes Estimation (Illustrative)

× - Least Squares Estimate

□ - Empirical Bayes Estimate, Shrinking to a Point

○ - Empirical Bayes Estimate, Shrinking to a Line

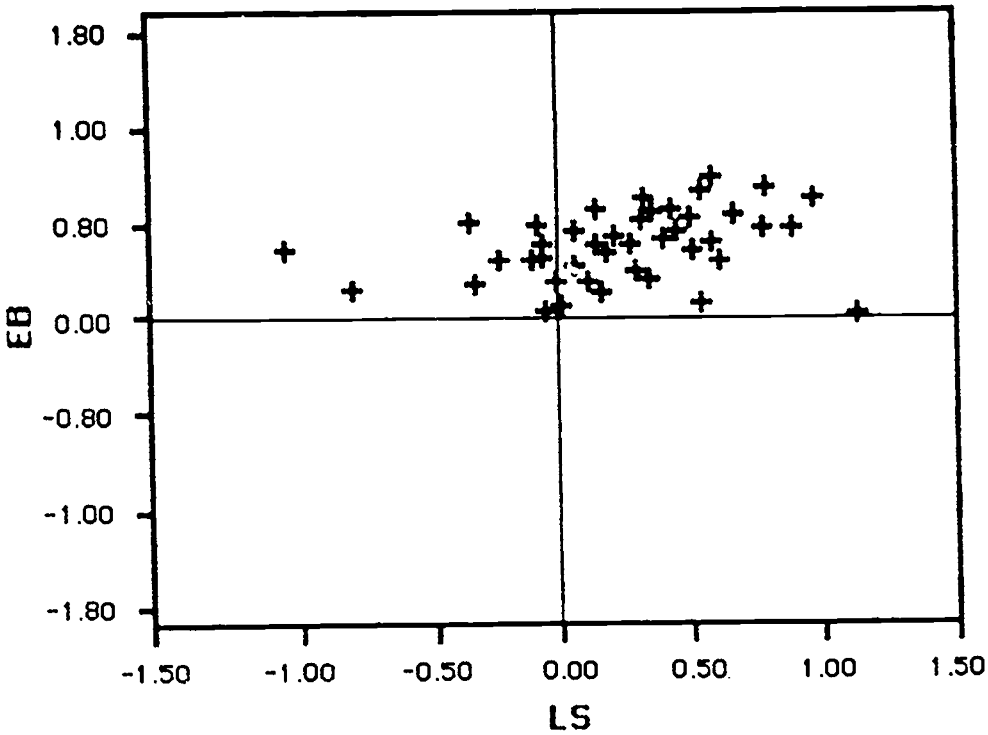

Figure 2: GRE DATA, EB VS. LS COEFFICIENTS, UGPA

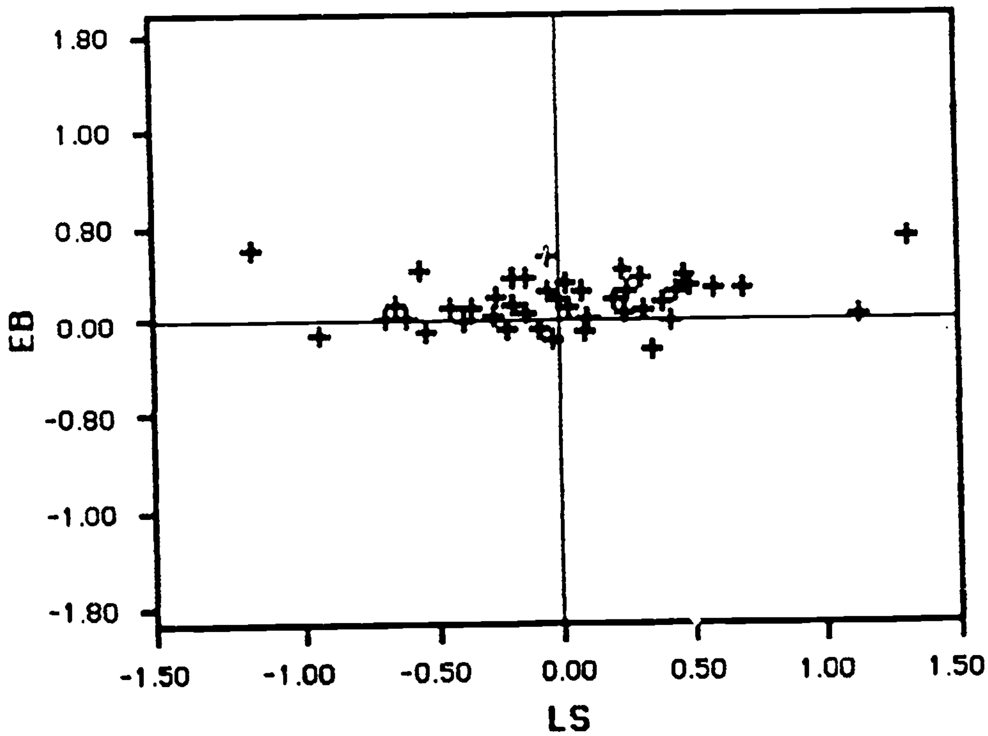

Figure 3: GRE DATA, EB VS. LS COEFFICIENTS, VERBAL GRE

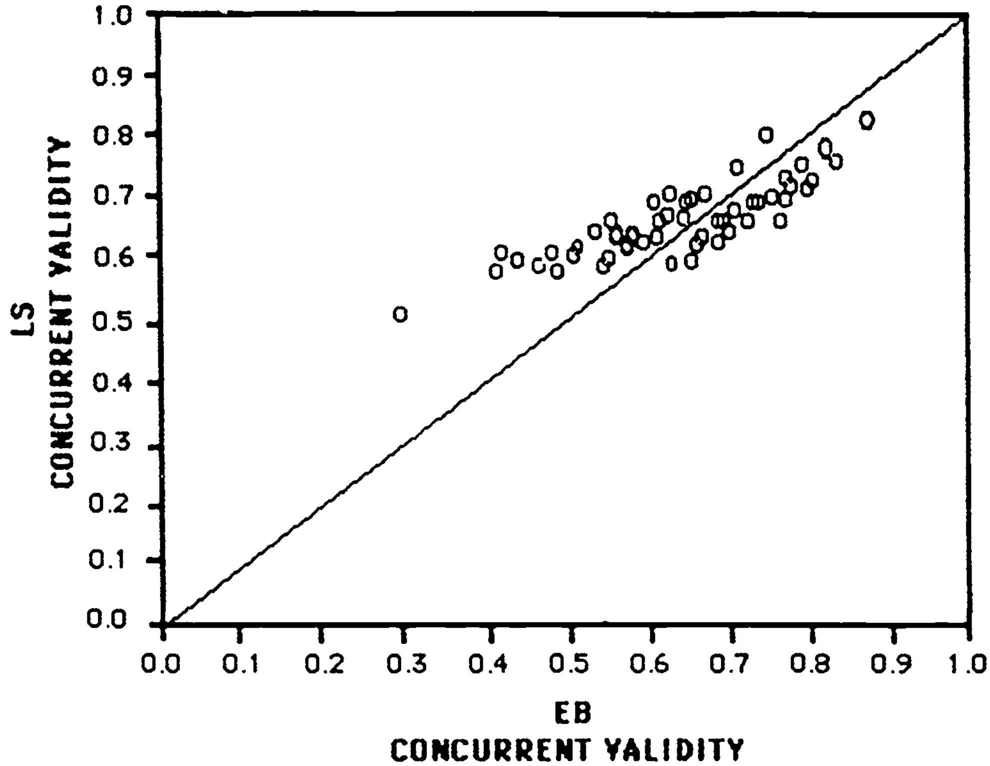

Figure 4: Empirical Bayes Concurrent Validity vs. Least Squares Concurrent Validity

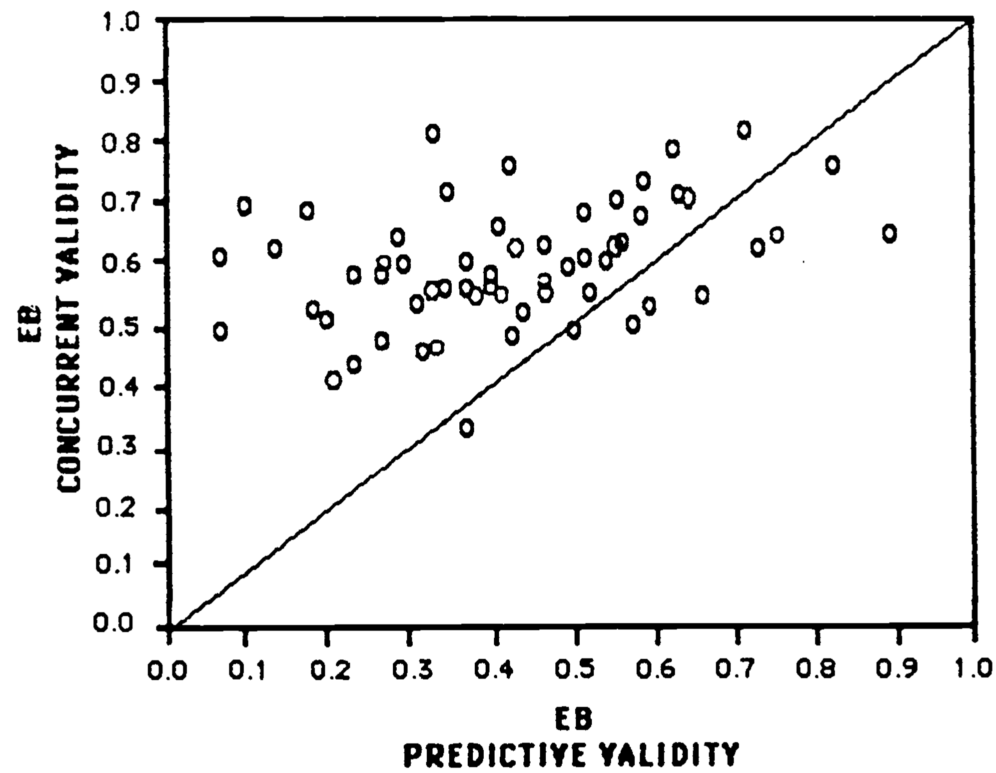

Figure 5: Empirical Bayes Predictive Validity vs. Empirical Bayes Concurrent Validity

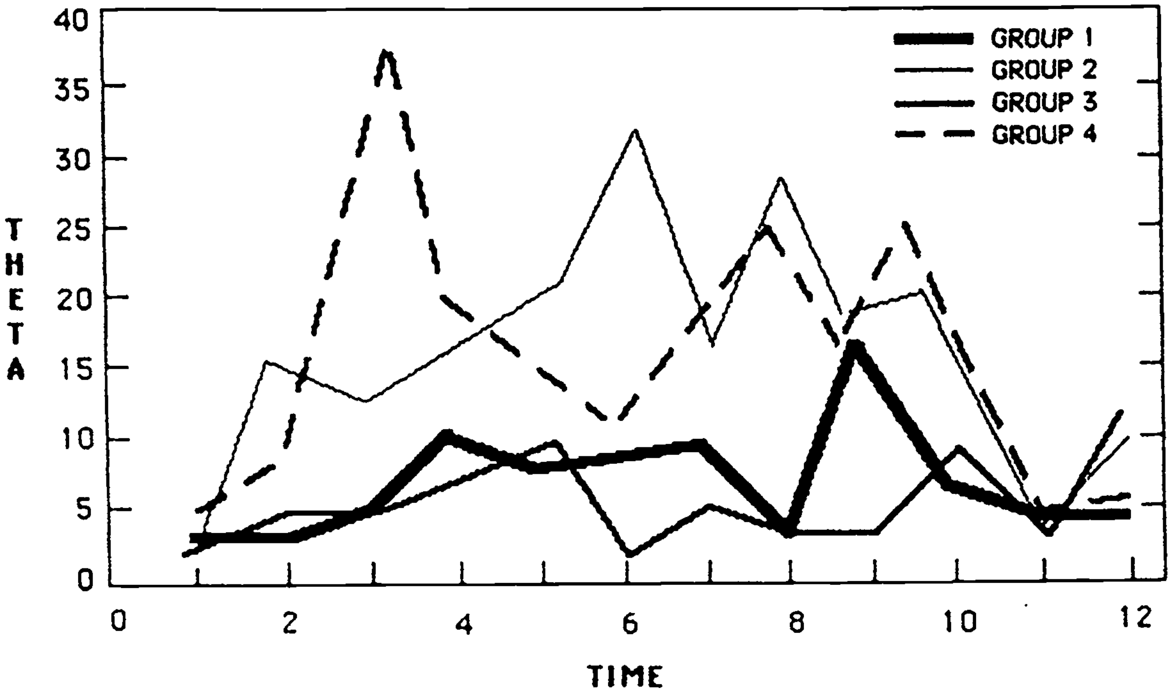

Figure 6: Occurrence/Exposure Rates. Cancer Data.

GROUP 1: Oophorectomy, Duke stage B
GROUP 2: Oophorectomy, Duke stage C
GROUP 3: No oophorectomy, Duke stage B
GROUP 4: No oophorectomy, Duke stage C

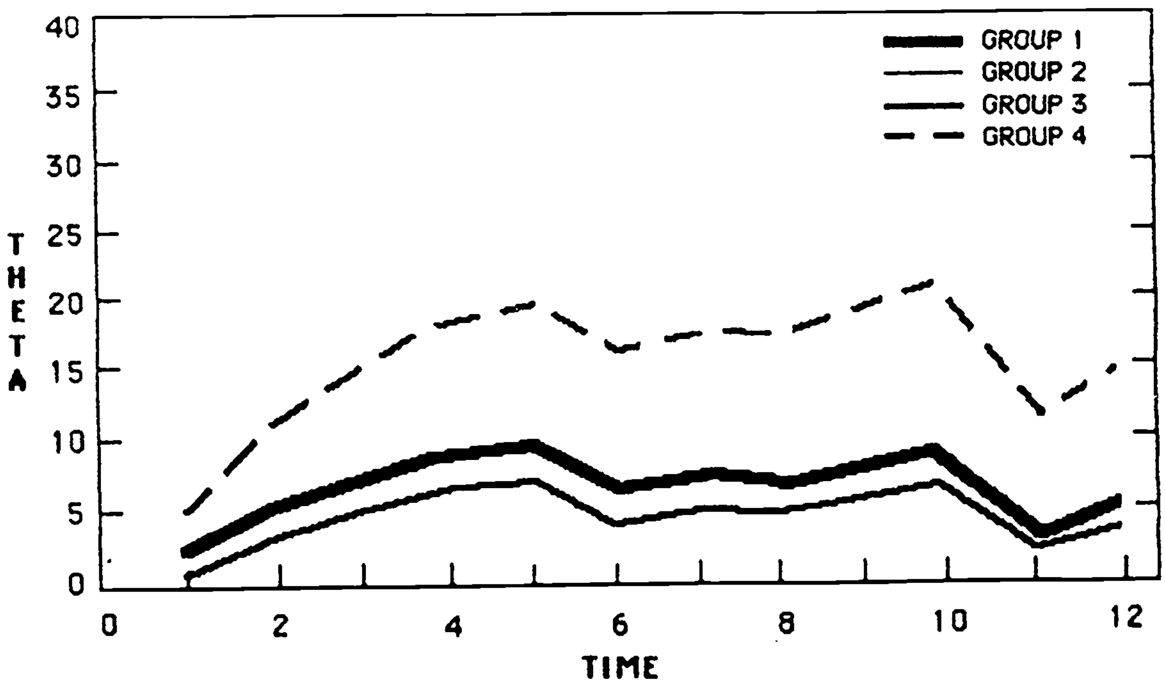

Figure 7: Empirical Bayes Estimates of Hazard Functions. Cancer Data.

References

Aitkin, M. & Clayton, D. (1980). The Fitting of Exponential Weibull and Extreme Value Distributions to Complex Censored Survival Data Using GLIM. Applied Statistics, 29, 156-163.

Berger, J. (1983). Discussion of paper by C. Morris. Journal of the American Statistical Association, 78, 55-57.

Braun, H. I., & Jones, D. H. (1981). The Graduate Management Admission Test: Prediction Bias Study. GMAC Research Report #81-4, Educational Testing Service, Princeton, NJ.

Braun, H. I., & Jones, D. H. (1985). Use of Empirical Bayes Methods in the Study of the Validity of Academic Predictors of Graduate School Performance. Research Report #84-34. Educational Testing Service, Princeton, NJ.

Braun, H. I., Jones, D. H., Rubin, D. B., & Thayer, D. T. (1983). Empirical Bayes Estimation of Coefficients in the General Linear Model from Data of Deficient Rank. Psychometrika, 48, 171-181.

Braun, H. I., Ragosta, M., & Kaplan, B. (1986a). The Predictive Validity of the Scholastic Aptitude Test for Disabled Students. Research Report #86-38, Educational Testing Service, Princeton, NJ.

Braun, H. I., Ragosta, M., & Kaplan, B. (1986b). The Predictive Validity of the Graduate Record Examination for Disabled Students. Research Report #86-42, Educational Testing Service, Princeton, NJ.

Braun, H. I., & Szatrowski, T. H. (1984). Validity Studies Based on a Universal Criterion Scale. Journal of Educational Statistics, 9, 331-344.

Cox, D. R. (1972). Regression Models with Life Tables (with Discussion). Journal of the Royal Statistical Society, Series B. 34, 187-220.

Cox, D. R. (1975). Partial Likelihood. Biometrika, 62, 269-276.

Cressie, N. (1982). A Useful Empirical Bayes Identity. Annals of Statistics, 10, 625-629.

Cutait, R., Lesser, M. L., & Enker, W. E. (1983). Prophylactic Oophorectomy in Surgery for Large Bowel Cancer. Diseases of the Colon and Rectum, 26, 6-11.

Deeley, J. J., & Lindley, D. V. (1981). Bayes Empirical Bayes. Journal of the American Statistical Association, 76, 833-841.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in Covariance Components Models. Journal of the American Statistical Association, 76, 341-353.

DerSimonian, R., & Laird, N. M. (1983). Evaluating the Effect of Coaching on SAT Scores: A Meta-Analysis. Harvard Educational Review, 53(1), 1-15.

Efron, B., & Morris, C. N. (1973). Stein's Estimation Rule and its Competitors - An Empirical Bayes Approach. Journal of the American Statistical Association, 68, 117-130.

Efron, B., & Morris, C. N. (1975). Data Analysis Using Stein's Estimator and its Generalizations. Journal of the American Statistical Association, 70, 311-319.

Gaver, D. P., & O'Muircheartaigh, I. G. (1987). Robust Empirical Bayes Analysis of Event Rates. Technometrics, 29, 1-15.

Glass, G. V. (1976). Primary, Secondary, and Meta-Analysis of Research. Educational Researcher, 5 3-8.

Grandell, J. (1972). Doubly Stochastic Poisson Processes. Institute of Actuarial Mathematics and Mathematical Statics, Stockholm: University of Stockholm.

Hedges, L. V., & Olkin, I. (1983). Regression Modes in Research Synthesis. American Statistician, 37(2), 137-140.

Hedges, L. V., & Olkin, I. (1985). Statistical Methods for Meta-Analysis. New York: Academic Press.

Hedges, L. V. (1987). The Meta-Analysis of Test Validity Studies: Some New Approaches. in Test Validity, (eds. H. Wainer & H. Braun, Hillsdale, NJ: Lawrence Erlbaum, Inc., in press.

Holford, T. R. (1980). The Analysis of Rates and Survivorship Using Log-Linear Models. Biometrics, 36, 299-305.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-Analysis: Cumulating Research Findings Across Studies, Beverly Hills: Sage.

James, W., & Stein, C. M. (1961). Estimation with Quadratic Loss. Procedures of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, 361-379, Berkeley: University of California Press.

Laird, N. M. (1978). Empirical Bayes Methods for Two-Way Contingency Tables. Biometrika, 65, 581-590.

Laird, N. M. (1982). Empirical Bayes Estimates Using the Nonparametric Maximum Likelihood Estimate for the Prior. Journal of Statistical Computation and Simulation, 15, 211-220.

Laird, N. M., Lange, N., & Stram, D. (1987). Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm. Journal of the American Statistical Association, 82, 97-105.

Laird, N. M., & Louis, T. (1982). Approximate Posterior Distributions for Incomplete Data Problems. Journal of the Royal Statistical Society, 44, 190-200.

Laird, N. M. & Oliver D. (1981). Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques. Journal of the American Statistical Association, 75, 231-240.

Leonard, T. (1983). Discussion of paper by C. Morris. Journal of the American Statistical Association, 78, 59-60.

Light, R. J., & Pillemer, D. B. (1984). Summing Up: The Science of Reviewing Research. Cambridge, MA: Harvard University Press.

Linn, R. L., Harnish, P. L., & Dunbar, S. B. (1981). Validity Generalization and Situational Specificity: An Analysis of the Prediction of First-Year Grades in Law School. Applied Psychological Measurement, 5(3), 281-289.

Linn, R. L., & Hastings, C. N. (1987). A Meta-Analysis of the Validity of Predictors of Performance in Law School. Journal of Educational Measurement, 21, 245-259.

Louis, T. A. (1984). Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods. Journal of the American Statisical Association. 79, 393-398.

Maritz, T. S. (1970). Empirical Bayes Methods. London: Methuen.

Mislevy, R. (1987). Exploiting Auxiliary Information About Items in the Estimation of Raush Item Difficulty Parameters. ETS Research Report, Princeton, NJ: to appear.

Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. Journal of the American Statistical Association, 78, 47-65 (with discussion).

Phadia, E. G. (1980). Nonparametric Bayesian Inference Based on Censored Data -- An Overview. In Nonparametric Statistical Inference, Vol. 32 of Colloquia Mathematica Societatis Janos Bolyai. Budapest.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes Meta-Analysis. Journal of Educational Statistics, 10, 75-98.

Rubin, D. B. (1980). Using Empirical Bayes Techniques in the Law School Validity Studies. Journal of the American Statistical Association, 75, 801-816.

Rubin, D. B. (1981). Estimation in Parallel Randomized Experiments. Journal of Educational Statistics, 6(4), 377-400.

Rosenthal, R., & Rubin, D. B. (1982). Comparing Effect Sizes of Independent Studies. Psychological Bulletin, 92, 500-504.

Schmidt, F. L. (1987). Validity Generalization and the Future of Criterion-related Validity. in Test Validity, (eds. Howard Wainer & Henry Braun), Hillsdale, NJ: Lawrence Erlbaum, Inc., in press.

Stone, M. (1978). Cross-validation: A Review. Mathematische Operationsforschung und Statistik, 9 127-140.

Swinton, S. (1986). The Predictive Validity of the Restructured GRE with Particular Attention to Older Students. Graduate Record Examination Final Report #83-25, Educational Testing Service, Princeton, NJ.

Szatrowski, T. H. (1976). Estimation and Testing for Block Compound Symmetry and Other Patterned Covariance Matrices with Linear and Non-Linear Structure. Technical Report #107, Stanford University, Stanford, CA.

Whitehead, J. (1980). Fitting Cox's Regression Model to Survival Data Using GLIM. Applied Statistics, 29, 268-275.

Wong, G. Y. & Mason, W. M. (1985). The Hierarchical Logistic Regression Model for Multilevel Analysis. Journal of the American Statistical Association, 80, 513-524.