The Annals of Statistics
1973, Vol. 1, No. 1, 93-103

AN EMPIRICAL BAYES APPROACH TO MULTIPLE LINEAR REGRESSION[1]

BY SERGE L. WIND

A. T. & T., Management Sciences Division
New York

We consider estimation (subject to quadratic loss) of the vector of coefficients of a multiple linear regression model in which the error vector is assumed to have 0 mean and covariance matrix $\sigma^2 I$ but is not assumed to take on a specific parametric form, e.g., Normal. The vector of coefficients is taken to be randomly distributed according to some unknown prior. Restricted minimax solutions are exhibited relative to equivalence classes on the space of all prior probability distributions which group distributions with the same specified moments. In the context of the classic Empirical Bayes formulation, we determine restricted asymptotically optimal estimators—i.e., decision functions whose Bayes risks converge to the risk of the restricted minimax decision at each component stage.

1. Introduction and summary. Estimators are exhibited for the parameters in the multiple linear regression model, which, in addition to using data from the present set of observations, incorporate data obtained in previous (independent) experiments. These estimators have lower Bayes risk than the usual least squares estimators.

The coefficient parameters are assumed to be random variables, distributed according to an unknown prior distribution. The regression problem is assumed to occur repeatedly and independently, with the same prior throughout—the Empirical Bayes formulation of Robbins [10]. The classic Robbins Empirical Bayes approach has been applied to the general linear model with Normally distributed error variables, e.g., see Martz and Krutchkoff [8], where consistent estimators for the (single stage) Bayes rule are proposed as Empirical Bayes procedures. In this paper, no specific parametric form is assumed, and the problem is formulated in the generalized Empirical Bayes approach first considered by Robbins in Section 5 of [10] as an application to estimation of a binomial parameter, and then expanded upon by Cogburn [3]. In this formulation, a restricted minimax decision, defined on an equivalence class, or subset of the set of priors, which groups distributions with the same specified moments, is the decision function whose Bayes risk is less than or equal to the infimum (over all decisions) of the supremum (over an equivalence class) of the Bayes risk function. We determine restricted asymptotically optimal estimators, i.e., decision functions

Received March 17, 1971; revised June 6, 1972.

[1] This work is a component of the research submitted for a Ph. D. degree at Columbia University, Dept. of Mathematical Statistics, and was supported in part by National Science Foundation Grant NSF-GP-9640.

AMS 1970 *subject classifications*. Primary 62C10, 62J05; Secondary 62G05.

Key words and phrases. Empirical Bayes, regression, Stein–James estimators, restricted minimax.

whose Bayes risks approach the risk of the restricted minimax decision at each component stage.

Satisfying the criterion of restricted asymptotic optimality are components of "Stein–James estimators." These estimators are generalizations and modifications of the one proposed by James and Stein [5] for estimating the mean of a multivariate Normal distribution, and they have been considered as Empirical Bayes procedures by Kantor [6] and Efron and Morris [4] for Normal distributions, and by Cogburn [3] for the case where no parametric distribution is specified. For the latter instance, properties of Stein–James estimators are given by Cogburn [2] and Wind [14].

Definitions of terms and a formalization of this approach in a statistical framework, based on concepts defined in [3], are presented in Section 2.

The regression model is analyzed in Section 3. It is assumed that $Y = X\beta + \varepsilon$, where X is a matrix of known constants, β is a vector of unknown parameters, and ε is a random vector with mean 0 and covariance $\sigma^2 I$. For quadratic loss, restricted minimax solutions are exhibited for classes of prior distributions on β (Theorem 3.1) and for joint priors on (β', σ^2) (Theorem 3.2). For equivalence relations on the priors specifying (i) their means and variances or (ii) second moments, restricted asymptotically optimal decision functions are found in Theorem 3.3 if each independent component regression problem has the same diagonal $X'X$ matrix and common unknown variance σ^2.

In Section 4, based on the work of Bhattacharya [1] and Sclove [11], a correspondence is developed between the problems of estimating regression coefficients and estimating location parameters. Restricted asymptotically optimal estimators of β relative to equivalences which group priors by their mean vectors and by all components of their covariance matrices are exhibited in Theorem 4.1.

2. Definitions and statistical framework. For the statistical decision problem with which we are concerned, the sample space, parameter space, decision space and their σ-fields are given respectively by $(\mathscr{X}, \mathscr{A})$, (Θ, τ), and $(\mathscr{E}, D)$. The family of distributions $\{F_\theta(\cdot): \theta \in \Theta\}$ satisfies $F_\cdot(B)$ τ-measurable for each $B \in \mathscr{A}$. The loss function L is $\tau \times D$ measurable. $R(\cdot, \delta)$ denotes the risk function of the randomized (behavioral) decision function δ. Δ is the space of all such decision procedures. $L^*(\theta, \delta, x) = \int_{\mathscr{E}} L(\theta, c)\delta(x, dc)$. Π denotes the set of all (prior) probability distributions on τ and $R^*(\pi, \delta) = \int_\Theta R(\theta, \delta)\pi(d\theta)$ is the Bayes risk function.

DEFINITION 2.1. The *Bayes envelope* function r^* defined on Π is $r^*(\pi) = \inf_{\delta \in \Delta} R^*(\pi, \delta)$.

DEFINITION 2.2. δ_π is *Bayes* re π if $R^*(\pi, \delta_\pi) = r^*(\pi)$.

DEFINITION 2.3. Let T be contained in the class of all subsets of Π. A decision function δ_T is *restricted minimax* relative to T if

$$R^*(\pi', \delta_T) \leqq \inf_\delta \sup_{\pi \in T} R^*(\pi, \delta), \qquad \forall \pi' \in T.$$

If a priori we know that $\pi \in T$, it is reasonable, in a minimax sense, to use a restricted minimax decision, if it exists. Thus, if we describe partial information about the prior distribution π by introducing an equivalence relation $\sim$ on Π, we assume we know π up to $\sim$. We will denote the equivalence class generated by π by

$$\sim(\pi) = \{\pi' \in \Pi : \pi' \sim \pi\} .$$

DEFINITION 2.4. The *envelope risk* function $r_\sim{}^*$, relative to $\sim$, is

$$r_\sim{}^*(\pi) = \inf_\delta \sup_{\sim\pi} R^*(\pi', \delta) ,$$

and is the best (in a minimax sense) we can do given that we know π up to $\sim$.

Let us now describe the Empirical Bayes problem. Given (Θ, τ), $(\mathscr{X}, \mathscr{A})$, and $(\mathscr{E}, D)$, measurable spaces, and loss function L as above. Let (θ_1, x_1), (θ_2, x_2), $\cdots$ be a sequence of pairs of random variables, each pair independent of the others, with the θ_i having a common a priori distribution $\pi \in \Pi$, and the distribution of x_i given $\theta_i = \theta$ denoted by $F_\theta(\cdot)$. So, at the nth stage or component experiment, we can say that a real vector $\theta^{(n)} = (\theta_1, \cdots, \theta_n)$ is selected by choosing the components independently, each with the same distribution π and then $x^{(n)} = (x_1, \cdots, x_n)$ has probability distribution $F_{\theta^{(n)}}(\cdot)$ with

$$F_{\theta^{(n)}}(x^{(n)}) = F_{\theta_1}(x_1) \cdot F_{\theta_2}(x_2) \cdot \cdots \cdot F_{\theta_n}(x_n) .$$

The parameter space is $\Theta^{(n)}$ and the experiment space $(\mathscr{X}^{(n)}, \mathscr{A}^{(n)})$, where the superscript denotes the n-fold Cartesian product. Let decision function δ_n be a transition probability mapping such that $\delta_n(x^{(n)}, \cdot)$ is a probability measure on the space $(\mathscr{E}, D)$ for each $x^{(n)} \in \mathscr{X}^{(n)}$, and $\delta_n(\cdot, c)$ is $\mathscr{A}^{(n)}$-measurable in $x^{(n)}$ for each $c \in D$; i.e., $\delta_n : \mathscr{X}^{(n)} \to \mathscr{E}$, whose form is functionally dependent on $x^{(n)}$. The overall expected loss of the nth stage estimator δ_n is

$$R_n{}^*(\pi, \delta_n) = \int_\Theta \pi(d\theta_1) \cdots \int_\Theta \pi(d\theta_n) \int_{\mathscr{X}^{(n)}} L_n{}^*(\theta_n, \delta_n, x^{(n)}) \cdot F_{\theta^{(n)}}(dx^{(n)}) ,$$

where

$$L_n{}^*(\theta_n, \delta_n, x^{(n)}) = \int_{\mathscr{E}} L(\theta_n, c)\delta_n(x^{(n)}, dc) .$$

DEFINITION 2.5. Given a sequence of decision functions $\{\delta_n\}$, $n = 1, 2, \cdots$ for the Empirical Bayes problem, if

$$R_n{}^*(\pi, \delta_n) \to \inf_{\delta_n} R_n{}^*(\pi, \delta_n)$$

for all $\pi \in \Pi$, then $\{\delta_n\}$ is called *asymptotically optimal*. As noted in (3.6) of [3],

$$\inf_{\delta_n} R_n{}^*(\pi, \delta_n) = \inf_\delta R^*(\pi, \delta) \equiv r^*(\pi) .$$

Suppose we select the equivalence relation that identifies those distributions of Π having the same specified set of moments, a characteristic of π which should be easily estimable from our $x^{(n)}$ sample. We hope the Bayes risks of our estimators do as well asymptotically as if the equivalence class were known beforehand and at each stage we used the corresponding restricted minimax decision. Thus this criterion, described in [3], becomes

$$R_n{}^*(\pi, \delta_n) \to r_\sim{}^*(\pi) \qquad \forall \pi . \tag{2.1}$$

DEFINITION 2.6. A sequence of estimators $\{\delta_n\}$ satisfying (2.1) is called *restricted asymptotically optimal.*

3. Linear regression. The multiple linear regression model consists of data

$$Z' = (Y, X),$$

where $y_j = \beta' x_j + \varepsilon_j$, $(j = 1, 2, \cdots, N)$, $x_j = (x_{j1}, \cdots, x_{jp})'$, with

$$E\varepsilon_j = 0 \quad \text{and} \quad E\varepsilon_j \varepsilon_k = \sigma^2 \delta_{jk}, \tag{3.1}$$

where δ_{jk} is the Kronecker delta. Or,

$$Y = X\beta + \varepsilon, \tag{3.2}$$

where

Y is an $N \times 1$ random observed vector,
ε is an $N \times 1$ random (error) vector with moments (3.1),
X is an $N \times p$ matrix of fixed quantities, and
β is a $p \times 1$ vector of unknown coefficients.

We will always assume that X is of full rank (rank $X = p < N$). Let $V = X'X$. Then the LSE is $\hat{\beta} = V^{-1}X'Y$.

Normality is not assumed in (3.1); a partially non-parametric approach is taken by specifying only the first two moments of the error variable.

The problem considered is point estimation of β. A general loss function, under which methods of this paper apply, is

$$\sigma^{-2}(\beta - b)'\Delta(\beta - b), \tag{3.3}$$

with Δ a known $p \times p$ positive definite matrix. However, the loss function

$$\sigma^{-2}(\beta - b)'V(\beta - b) \tag{3.4}$$

will be used frequently, since it can be considered "natural" in the following sense. Suppose we have another set of data Z^*, independent of Z, but with the same distribution; i.e., $y_k^* = \beta' x_k^* + \varepsilon_k$, $k = 1, \cdots, N$ with moments (3.1). We are given X^* and wish to predict Y^*, by taking an estimator b of β which is a function only of Z. Then the mean squared error of prediction, conditional on b, appropriately transformed so that the loss of a perfect estimate is zero, can be shown, using the same proof as given by Stein [12] for the Normal linear regression model with random predictors, to be precisely (3.4).

Groups of prior distributions on the p-vector β are identified by specifying a set of moments of the prior. Assuming variance σ^2 is known, the moments considered that determine an equivalence class are the p means and the $p(p + 1)/2$ elements of the covariance matrix. Each class must be identified by at least p moments—the second moments of the marginals—and can be identified by as many as $(p^2 + 3p)/2$ moments, including the means and covariances.

Let $\mu = (\mu_1, \cdots, \mu_p)'$ and let Σ^* be a $p \times p$ symmetric matrix with components σ_{ij}^*. For the ith regression coefficient, if only the second moment is specified in

a given equivalence class, set $\sigma_{ii}^* = E\beta_i^2$ and $\mu_i = 0$; if the variance and mean are specified, set $\sigma_{ii}^* = \mathrm{Var}(\beta_i)$ and $\mu_i = E\beta_i$. For $j \neq i$, let $\sigma_{ij}^* = \mathrm{Cov}(\beta_i, \beta_j)$ if the covariance and the means of β_i and β_j are specified in the class; let $\sigma_{ij}^* = E\beta_i\beta_j$ if that moment is part of the equivalence and both the means of β_i and β_j are not part of the equivalence specification; let $\sigma_{ij}^* = 0$ if a covariance is not specified in the class.

Restricted minimax decisions for the (single stage) regression model are given below in Theorems 3.1 (for priors on β) and 3.2 (for joint priors on (β', σ^2)). First, two useful lemmas are stated.

LEMMA 3.1. *The Bayes estimator δ of β for generalized quadratic loss* (3.3) *and for specified parametric distributions on ε and β is $\delta = E(\beta \mid \hat{\beta})$, the mean of the posterior distribution of β given $\hat{\beta}$, and*

$$R^*(\pi, \delta) = \sigma^{-2} \sum_{i=1}^{p} \sum_{j=1}^{p} d_{ij} E[\mathrm{Cov}(\beta_i, \beta_j \mid \hat{\beta})] ,$$

where d_{ij} are the components of Δ and where the expectation is taken with respect to the marginal density of $\hat{\beta}$. (See, e.g., [7] pages 18–20.)

LEMMA 3.2. *If δ is Bayes with respect to some $\pi^* \sim \pi$ and $R^*(\pi', \delta)$ is constant for $\sim (\pi)$, then δ is restricted minimax relative to $\sim (\pi)$.*

THEOREM 3.1. *With loss* (3.4), *for the regression model* (3.2) *with σ^2 known, relative to a given equivalence class of the priors on β, the restricted minimax estimator is*

$$b = (\sigma^2(\Sigma^*)^{-1} + V)^{-1}(\sigma^2(\Sigma^*)^{-1}\mu + V\hat{\beta}) , \tag{3.5}$$

and $r_{\sim}^*(\pi) = \sum_i \sum_j v_{ij} m_{ij}$, where m_{ij} are the components of $M = (\sigma^2(\Sigma^*)^{-1} + V)^{-1}$.

PROOF. Consider the parametric subfamilies of Normal distributions of ε and β. The posterior distribution of β given LSE $\hat{\beta}$ is Normal with mean (3.5), which is the Bayes estimator, and covariance $\sigma^2 M$ ([9] page 337). Since its risk is constant for the equivalence class, b is restricted minimax for the Normal subfamily and thus for all distributions with first two moments specified.

If σ^2 is a random variable, an equivalence class of the joint priors on (β', σ^2) must be identified by at least $p + 1$ moments (the second moments on β and $E\sigma^2$) and by at most $(p^2 + 3p + 2)/2$.

THEOREM 3.2. *For the regression model* (3.2), *subject to loss $(b - \beta)'V(b - \beta)$, and with $\rho = E\sigma^2$ as part of the specification of the equivalence class of the priors on (β', σ^2), the restricted minimax estimator is*

$$b^* = (\rho(\Sigma^*)^{-1} + V)^{-1}(\rho(\Sigma^*)^{-1}\mu + V\hat{\beta}) , \tag{3.6}$$

and

$$r_{\sim}^*(\pi) = \rho \sum_i \sum_j v_{ij} l_{ij} , \qquad \text{where} \quad L = (\rho(\Sigma^*)^{-1} + V)^{-1} .$$

PROOF. If ε is Normally distributed and the prior on (β', σ^{-2}) is a Normal-gamma ([9] pages 343–345) with parameters $(\mu, t, \rho(\Sigma^*)^{-1}, \nu)$ with ν, t such that $\rho = \nu t/(\nu - 2)$ and rank $(\Sigma^*)^{-1} = p_1$, the marginal posterior distribution of β given

$\hat{\beta}$ is a multivariate Student distribution ([9] pages 256–259) with mean b^* (3.6), which is constant risk Bayes, and covariance $L\nu^*t^*/(\nu^* - 2)$, where

$$\nu^* = \nu + p_1 + N - p_2 ,$$
$$p_2 = \text{rank}\,(L^{-1}) ,$$
$$t^* = (\rho(\nu - 2) + \rho\mu'(\Sigma^*)^{-1}\mu + Y'Y - b^{*\prime}L^{-1}b^*)/\nu^* .$$

But $E(\nu^*t^*/(\nu^* - 2)) = \nu t/(\nu - 2) = \rho$, with the expectation taken with respect to $\hat{\beta}$.

Consider (3.2) with orthogonal design—i.e., $V = X'X$ is diagonal—and those equivalence classes which identify prior distributions on β having the same specified finite means and variances or the same second moments of β (but with no specification of covariance terms). A typical equivalence relation in this subgroup considered is $(\lambda_1, \cdots, \lambda_p)$, where λ_i is equal either to (μ_i, Ψ_i^2) or γ_i for $i = 1, \cdots, p$, and where $\mu_i = E\beta_i$, $\Psi_i^2 = \text{Var}(\beta_i)$, and $\gamma_i = E\beta_i^2$. The corollary below follows as a special case of Theorem 3.1 with diagonal Σ^* and V.

COROLLARY 3.1. *Given the regression model* (3.2) *with orthogonal design,* $b = (b_1, \cdots, b_p)'$ *is restricted minimax relative to* $(\lambda_1, \cdots, \lambda_p)$, *subject to loss* (3.4), *where*

$$b_i = \frac{\Psi_i^2\hat{\beta}_i + \sigma^2 v^{ii}\mu_i}{\sigma^2 v^{ii} + \Psi_i^2} \qquad \text{if} \quad \lambda_i = (\mu_i, \Psi_i^2) \tag{3.7}$$
$$= \frac{\gamma_i\hat{\beta}_i}{\sigma^2 v^{ii} + \gamma_i} \qquad \text{if} \quad \lambda_i = \gamma_i$$

for $i = 1, \cdots, p$, *and* $R^*(\pi, b) = \Sigma\phi_i(\phi_i + \sigma^2 v^{ii})^{-1}$, *where*

$$\phi_i = \Psi_i^2 \qquad \text{if} \quad \lambda_i = (\mu_i, \Psi_i^2)$$
$$= \gamma_i \qquad \text{if} \quad \lambda_i = \gamma_i .$$

PROPERTY 3.1. For estimating β_i, with marginal prior π_i, and letting b_{i1} denote the restricted minimax solution (3.7) for $\lambda_i = (\mu_i, \Psi_i^2)$, and b_{i2} denote (3.7) for $\lambda_i = \gamma_i$, then

$$R^*(\pi_i, b_{i1}) = \frac{\Psi_i^2}{\Psi_i^2 + \sigma^2 v^{ii}} \leqq \frac{\gamma_i}{\gamma_i + \sigma^2 v^{ii}} = R^*(\pi_i, b_{i2}) ,$$

with equality only if $\mu_i = 0$. Both risks are less than one, the risk of LSE $\hat{\beta}_i$.

Thus far, the treatment has been Bayesian. However, suppose the parameters of the prior are unknown, but we have available data from previous regression experiments—the "q-stage Empirical Bayes regression problem": Given a sequence of independent regression problems with data $Z_1, Z_2, \cdots, Z_q$ and, respectively, regression coefficients $\beta_{(1)}, \beta_{(2)}, \cdots, \beta_{(q)}$, where $Z_i = \{\binom{y_1 i}{x_1}\binom{y_2 i}{x_2} \cdots \binom{y_N i}{x_N}\}$ is the data of the ith stage, where $x_k = (x_{k1}, \cdots, x_{kp})'$ and $\beta_{(i)} = (\beta_{1i}, \cdots \beta_{pi})'$ and

$$y_{kj} = \textstyle\sum_{i=1}^{p} \beta_{ij} x_{ki} + \varepsilon_{kj} = \beta'_{(j)} x_k + \varepsilon_{kj} , \tag{3.8}$$

where

y_{kj} is the kth observation ($k = 1, \cdots, N$) of the jth component stage ($j = 1, \cdots, q$);
β_{ij} is the ith parameter ($i = 1, \cdots, p$) of the jth stage;
x_{ki} is the independent variable corresponding to the kth observation and ith regression coefficient;
ε_{kj} has moments (3.1) and ε_{ki}, ε_{cj} are independent for $j \neq i$, for all k and c.

Each experiment has the same independent variables, the same number of observations N, and common variance σ^2. $(\beta_{(1)}, Z_1), (\beta_{(2)}, Z_2), \cdots$ is a sequence of pairs of sets of random variables, each pair independent of the others. $\beta_{(1)}, \beta_{(2)}, \cdots$ is a sequence of independent realizations of random variables, each distributed according to the same (unknown) prior distribution.

Set

$$\hat{\sigma}^2 = \frac{1}{q(N-p)+2} \sum_{j=1}^{q} \sum_{i=1}^{N} (y_{ij} - \sum_{k=1}^{p} \hat{\beta}_{kj} x_{ik})^2 , \tag{3.9}$$

$$B_{iq}(\hat{\beta}_{iq}) = \left(1 - \frac{(q-1)\hat{\sigma}^2 v^{ii}}{\sum_j (\hat{\beta}_{ij} - \bar{\hat{\beta}}_i)^2}\right)^+ (\hat{\beta}_{iq} - \bar{\hat{\beta}}_i) + \bar{\hat{\beta}}_i , \tag{3.10}$$

$$D_{iq}(\hat{\beta}_{iq}) = \left(1 - \frac{q\hat{\sigma}^2 v^{ii}}{\sum_j \hat{\beta}_{ij}^2}\right)^+ \hat{\beta}_{iq} , \tag{3.11}$$

with

$$\bar{\hat{\beta}}_i = q^{-1} \sum_{j=1}^{q} \hat{\beta}_{ij} , \tag{3.12}$$

where $\hat{\beta}_{ij}$ is the jth stage LSE of β_{ij}, and $Z^+ \equiv \max(0, Z)$.

THEOREM 3.3. *Given the q-stage Empirical Bayes regression problem with orthogonal design, with the error variables ε drawn from a family of uniformly square integrable distributions with bounded fourth moments, the sequence of restricted asymptotically optimal decision functions relative to $(\lambda_1, \cdots, \lambda_p)$ for estimating $\beta_{1q}, \cdots, \beta_{pq}$ is $\{\hat{b}_{1q}, \cdots, \hat{b}_{pq}\}$, $q = 1, \cdots$ where $\lambda_i = (\mu_i, \Psi_i^2)$ or γ_i and*

$$\begin{aligned} \hat{b}_{iq} &= B_{iq} \quad \text{if} \quad \lambda_i = (\mu_i, \Psi_i^2) \quad \text{and} \quad q \geqq 3 \\ &= D_{iq} \quad \text{if} \quad \lambda_i = \gamma_i \quad \text{and} \quad q \geqq 3 \\ &= \hat{\beta}_{iq} \quad \text{if} \quad q = 1, 2 \end{aligned} \tag{3.13}$$

with B_{iq} defined in (3.10) *and D_{iq} in* (3.11).

PROOF. Follows from a general result given as Corollary 2.3 in [14], where Stein–James estimators of this form are considered. (See also [13].)

PROPERTY 3.2. The form of the Stein–James estimators (3.10) and (3.11) is suggested by substituting in (3.7) for the now unknown parameters μ_i, Ψ_i^2, γ_i,

and σ^2, estimators $\hat{\mu}_i = \bar{\hat{\beta}}_i$ (3.12), $\hat{\Psi}_i^2$, $\hat{\gamma}_i$, and $\hat{\sigma}^2$ (3.9) respectively, with

$$\hat{\Psi}_i^2 = [\sum_{j=1}^q (\hat{\beta}_{ij} - \bar{\hat{\beta}}_i)^2 - (q-1)\hat{\sigma}^2 v^{ii}]^+/(q-1)\,,$$
$$\hat{\gamma}_i = [\sum_{j=1}^q \hat{\beta}_{ij}^2 - q\hat{\sigma}^2 v^{ii}]^+/q\,.$$

THEOREM 3.4. *Under the conditions of Theorem* 3.3, *if the variance of* ε *varies with each component problem and a joint prior on* (β', σ^2) *is assumed, then relative to* $(\beta - b)'V(\beta - b)$, (3.13) *is restricted asymptotically optimal with* $\hat{\sigma}^2$ (3.9) *replaced by*

$$\bar{\hat{\sigma}}^2 = \frac{1}{q(N-p+2)} \sum_i^N \sum_j^q (y_{ij} - \sum_{k=1}^p \hat{\beta}_{kj} x_{ik})^2\,. \tag{3.14}$$

PROOF. Follows from Corollary 2.4 of [14].

PROPERTY 3.3. The results of the above two Theorems hold if in (3.8) the restrictions that the number of observations N and the independent observations X remain the same for each experiment are relaxed subject to the condition that $X'X$ remains constant for each component problem. Then, in place of (3.9), we estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{\sum_{j=1}^q (N_j - p) + 2} \sum_{j=1}^q \sum_{k=1}^{N_j} (y_{kj} - \sum_{i=1}^p \hat{\beta}_{ij} x_{kij})^2\,,$$

where x_{kij} denotes the kth observation ($k = 1, \cdots, N_j$) of the ith parameter ($i = 1, \cdots, p$) at the jth stage ($j = 1, \cdots, q$), and N_j is the number of observations at stage j.

PROPERTY 3.4. The form of the estimator B_{iq} (3.10) suggests that we are making a preliminary test of the hypothesis H_1: $\beta_{ij} = \bar{\beta}_i$ for all $j = 1, \cdots, q$. We accept H_1 if $\sum_j (\hat{\beta}_{ij} - \bar{\hat{\beta}}_i)^2 \leqq (q-1)\hat{\sigma}^2 v^{ii}$, and then estimate β_{ij} by $\bar{\hat{\beta}}_i$. If we accept H_1, we are really pooling independent consistent estimators $\hat{\beta}_{i1}, \cdots, \hat{\beta}_{iq}$ and using $\bar{\hat{\beta}}_i$.

4. Restricted asymptotically optimal solutions. In this section, a correspondence between estimating regression coefficients in the q-stage Empirical Bayes formulation and estimating mean vectors also in a q-stage Empirical Bayes setup is established. This correspondence is employed in Theorem 4.1 to present restricted asymptotically optimal decision functions for the regression problem.

Given, as specified in (3.8), a sequence of q independent regression problems with data $Z_1, \cdots, Z_q$, we are interested in estimating the $p \times q$ matrix $\boldsymbol{\beta}$. For convenience in handling the data, (3.8) is rewritten in the following form: Observe a sample of N observations $Y_1, \cdots, Y_N$ on the dependent variables, where the q-dimensional vector $Y_k = (y_{k1}, \cdots, y_{kq})'$ is distributed with mean $\boldsymbol{\beta}' x_k$ and nonsingular covariance matrix $\Sigma = \sigma^2 I$. Set the $q \times N$ matrix $Y = (Y_1, \cdots, Y_N)$ and the $p \times N$ matrix $X' = (x_1, \cdots, x_N)$. Then, conditional on $\boldsymbol{\beta}$, $EY = \boldsymbol{\beta}' X'$. The loss function is

$$L_1(\boldsymbol{\beta}, \mathbf{b}) = \operatorname{tr} \sigma^{-2}(\mathbf{b} - \boldsymbol{\beta})'\Delta(\mathbf{b} - \boldsymbol{\beta})\,, \tag{4.1}$$

the q-variate analog to (3.3), with tr denoting the trace.

The least squares estimate of β is $\hat{\beta} = V^{-1}U$, with $U = \sum_{k=1}^{N} x_k Y_k'$, $V = X'X = \sum_{k=1}^{N} x_k x_k'$. Estimate σ^2 by (3.9), or identically, set

$$\hat{\sigma}^2 = (q(N-p) + 2)^{-1} \operatorname{tr} S , \tag{4.2}$$

with

$$S = T - U'V^{-1}U = T - \hat{\beta}' V \hat{\beta} , \qquad T = \sum_{k=1}^{N} Y_k Y_k' .$$

The transformations on the regression problem given below were suggested by the mappings in [1] and [11] which assumed a Normal process, but the imposition here of a prior distribution necessitated some modifications.

There exists a nonsingular $p \times p$ matrix C such that

$$C'C = V \quad \text{and} \quad C'DC = \Delta . \tag{4.3}$$

Let $W^{(1)} = C'^{-1}X'$, and the rows of $W^{(1)}$ are orthogonal. Following Sclove [11], select the $(N-p) \times N$ matrix $W^{(2)}$ such that $W' = (W^{(1)\prime}, W^{(2)\prime})$ is an $N \times N$ orthogonal matrix. Set $Z^* = YW'$. Then $Z^* = (Z^{(1)}, Z^{(2)}) = Y(W^{(1)\prime}, W^{(2)\prime})$, where $Z^{(1)}$ is of dimension $q \times p$ and $Z^{(2)}$ is $q \times (N-p)$. Let $Z_i = (Z_{1i}, \cdots, Z_{qi})'$ be a q-vector for $i = 1, \cdots, N$. Then $Z^* = (Z_1, \cdots, Z_p, Z_{p+1}, \cdots, Z_N)$. It is easy to verify that $\hat{\beta} = C^{-1}Z^{(1)\prime}$, $EZ^{(1)} = (\mathbf{E}\beta)'C'$,

$$S = Z^{(2)}Z^{(2)\prime} = \sum_{k=p+1}^{N} Z_k Z_k' ,$$

and $EZ^{(2)} = 0$, with $\mathbf{E}\beta$ the mean of β. Set the $q \times p$ matrix $\theta = (\theta_1, \cdots, \theta_p) = \beta'C' = E(Z^{(1)} \mid \beta)$. Then, for $i = 1, \cdots, p$, $Z_i = \hat{\beta}'(C')_i$ has mean θ_i and covariance $\Sigma = \sigma^2 I$, conditional on β, where $(C')_i$ denotes the ith column of C'. For $i = p+1, \cdots, N$, Z_i has mean 0 and covariance Σ. Conditional on β, Z_i, Z_j are mutually orthogonal and $Z^{(1)}$ and $Z^{(2)}$ lie in orthogonal spaces.

Let P be a $p \times p$ orthogonal matrix which rotates the axis of the distribution such that for each row of the transformed variable

$$K = Z^{(1)}P' , \tag{4.4}$$

the elements are pairwise uncorrelated. Recall $\Sigma^* = \operatorname{Cov}(\beta)$.

Thus $P(\sigma^2 I + C\Sigma^* C')P'$ is the diagonal covariance matrix for the data of each component experiment. (P is the matrix such that the elements of $P(Z^{(1)\prime})_i$, for each $i = 1, \cdots, q$, are uncorrelated unconditionally where $(Z^{(1)\prime})_i$ is the ith column of $Z^{(1)\prime}$.) Let $K = (K_1, \cdots, K_p)$, each K_i a q-vector. K_i, K_j $(j \neq i)$ are unconditionally uncorrelated and the components of each K_i are mutually independent.

Under transformations (4.3) and (4.4), $K = \hat{\beta}'C^{*\prime} = YW^{(1)\prime}P'$ with $C^* = PC$ and $C^{*\prime}C^* = V$. Let $\hat{\theta}$ be an estimator of $\theta = \beta'C'$ and let $\hat{\eta}$ be an estimator of $\eta = \beta'C^{*\prime}$; define the corresponding estimator $\mathbf{b}$ of β as $\mathbf{b} = C^{*-1}\hat{\eta}'$.

PROPERTY 4.1. For estimating η, loss (4.1) is

$$L_2(\eta, \hat{\eta}) = \sigma^{-2} \sum_{i=1}^{p} d_i(\hat{\eta}_i - \eta_i)'(\hat{\eta}_i - \eta_i) = L_1(\beta, \mathbf{b}) . \tag{4.5}$$

We can now see how the problem of estimating β with loss (4.1) is related to

the problem of estimating mean η subject to (4.5) with data K, with uncorrelated column vectors $K_1, \cdots, K_p$, with components of each K_i independent, and data S, which lies in a space orthogonal to K. For an estimator $\varphi(K_1, \cdots, K_p, S)$ of η, define the corresponding estimator for $\boldsymbol{\beta}$,

$$\Psi(\hat{\boldsymbol{\beta}}, S) = C^{*-1}\varphi'(\hat{\boldsymbol{\beta}}'(C^{*\prime})_1, \cdots, \hat{\boldsymbol{\beta}}'(C^{*\prime})_p, S) . \tag{4.6}$$

With respect to losses (4.1) and (4.5), let

$$R_1(\boldsymbol{\beta}, \Psi) = EL_1(\boldsymbol{\beta}, \Psi) \quad \text{and} \quad R_2(\eta, \varphi) = EL_2(\eta, \varphi) .$$

PROPERTY 4.2. $$R_2(\eta, \varphi) = R_1(\boldsymbol{\beta}, \Psi) .$$

It follows that (4.6) preserves dominance relations among estimators, and the theorem below can be verified.

THEOREM 4.1. *Given the q-stage Empirical Bayes regression problem* (3.8) *with the error variables* ε *drawn from a family of uniformly square integrable distributions with bounded fourth moments, the restricted asymptotically optimal estimator at stage* q, *relative to a given equivalence class of priors of* β *specified by its first two moments and loss* (3.4), *is*

$$C^{*-1}(\delta_{q1}(K_{q1}), \cdots, \delta_{qp}(K_{qp}))' ,$$

with $K = \hat{\boldsymbol{\beta}}'C^{*\prime}$, *and for* $i = 1, \cdots, p$,

$$\begin{aligned}
\delta_{qi}(K_{qi}) &= \delta_i(K_{qi}) \quad \text{if} \quad q \geqq 3 \quad \text{and} \quad \rho_i \neq 0 \\
&= \delta_i{}^*(K_{qi}) \quad \text{if} \quad q \geqq 3 \quad \text{and} \quad \rho_i = 0 \\
&= \bar{K}_i \quad \text{if} \quad q < 3 ,
\end{aligned}$$

$$\delta_i(K_{qi}) = \left(1 - \frac{(q-1)\hat{\sigma}^2}{\sum (K_{ji} - \bar{K}_i)^2}\right)^+ (K_{qi} - \bar{K}_i) + \bar{K}_i ,$$

$$\delta_i{}^*(K_{qi}) = \left(1 - \frac{q\hat{\sigma}^2}{\sum_j K_{ji}^2}\right)^+ K_{qi} ,$$

$$\bar{K}_i = q^{-1} \textstyle\sum_{j=1}^{q} K_{ji} ,$$

$$\rho_i = \textstyle\sum_{d=1}^{p} c_{id}^{*} \mu_d ,$$

with μ_d *as specified prior to Theorem* 3.1 *and* $\hat{\sigma}^2$ *given by* (4.2).

PROPERTY 4.3. If the component problems do not have common variance σ^2, let Σ be a diagonal matrix with elements $\sigma_1^2, \cdots, \sigma_q^2$. For a joint prior on (β', σ^2), replace $\hat{\sigma}^2$ in Theorem 4.1 with (3.14), and the results obtain subject to loss $(b - \beta)'\Delta(b - \beta)$.

Acknowledgment. I wish to express my appreciation to Professors Herbert Robbins and A. J. Baranchik for their encouragement and guidance of my Ph. D. dissertation research, of which this work is a part.

REFERENCES

[1] BHATTACHARYA, P. K. (1966). Estimating the mean of a multivariate normal population with general quadratic loss function. *Ann. Math. Statist.* **37** 1819–1824.

[2] COGBURN, R. (1965). On the estimation of a multivariate location parameter with squared error loss, in *Bernoulli (1723), Bayes (1763) and Laplace (1813) Anniversary Volume*, ed. J. Neyman and L. LeCam. Springer-Verlag, Berlin, 24–29.

[3] COGBURN, R. (1967). Stringent solutions to statistical decision problems. *Ann. Math. Statist.* **38** 447–463.

[4] EFRON, B. and MORRIS, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators—Part II: The empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139.

[5] JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1**. Univ. of California Press, 361–379.

[6] KANTOR, M. (1967). Estimating the mean of a multivariate normal distribution with applications to time series and empirical Bayes estimation. Ph. D. dissertation, Columbia Univ.

[7] MARTZ, H. (1967). Empirical Bayes estimation in multiple linear regression. Ph. D. dissertation, Virginia Polytechnic Institute.

[8] MARTZ, H. and KRUTCHKOFF, R. (1969). Empirical Bayes estimators in a multiple linear regression model. *Biometrika* **56** 367–374.

[9] RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*. The M.I.T. Press.

[10] ROBBINS, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35** 1–20.

[11] SCLOVE, S. (1967). Improved estimation of regression parameters. Technical Report No. 125, Dept. of Statistics, Stanford Univ.

[12] STEIN, C. (1960). Multiple regression, in *Contributions to Probability and Statistics—Essays in Honor of Harold Hotelling*. Stanford Univ. Press, 424–443.

[13] WIND, S. (1970). An empirical Bayes approach to the multiple linear regression problem. Ph. D. dissertation, Columbia Univ.

[14] WIND, S. (1972). Stein-James estimators of a multivariate location parameter. *Ann. Math. Statist.* **43** 340–343.

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
MANAGEMENT SCIENCES DIVISION
130 JOHN STREET
ROOM 1761
NEW YORK, NEW YORK 10038