The Annals of Statistics
1981, Vol. 9, No. 5, 960–973

THE ADMISSIBLE BAYES CHARACTER OF SUBSET SELECTION TECHNIQUES INVOLVED IN VARIABLE SELECTION, OUTLIER DETECTION, AND SLIPPAGE PROBLEMS

BY RONALD W. BUTLER

The University of Texas

We demonstrate the admissible Bayes character of two residual error criteria for subset selection of independent variables in normal multivariate regression models. In particular, suppose a linear model includes the independent variable list X, and suppose all additional independent variable subsets of size s (fixed) from list Z are under consideration for inclusion in the model. Let $\mathbf{\Sigma}$ be the regression error covariance matrix and $\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}_{\mathscr{S}})$ the usual unbiased estimator of Σ which assumes a model fitting the variables in list $X(X \cup \mathscr{S}$, where $\mathscr{S} \subset Z$ of size s). Then two *best* subsets of Z of size s may be characterized as minimizing tr $\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}}_{\mathscr{S}}$ and $|\hat{\mathbf{\Sigma}}_{\mathscr{S}}|/|\hat{\mathbf{\Sigma}}|$ over all subsets $\mathscr{S}$ of size s in Z. We show that the significance tests for including-excluding these two best subsets are admissible proper Bayes rules for fixed effect variables in list Z. If Z is allowed to encompass random and mixed effects, then the latter test is admissible proper Bayes in a class of location and scale invariant tests. Special cases of the general selection problem include multiple outlier detection and slippage tests where the best subset criteria above lead to *Studentized residual* outlier and slippage detection criteria. These tests are derived using models which explain outliers and slippage as locational biases and/or inflated variances.

1. Introduction. This paper demonstrates the admissible Bayes character of two selection criteria for independent variables in normal multivariate regression models. In particular, suppose a linear model includes the independent variable matrix $\mathbf{X}$. Let the columns of $n \times m$ matrix $\mathbf{Z}$ represent m additional independent variables that are under consideration for inclusion in the model. Suppose $\mathbf{\Sigma}$ is the regression error covariance matrix, and $\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}_{\mathscr{S}})$ is the usual unbiased estimator of $\mathbf{\Sigma}$ which assumes a model fitting $\mathbf{X}$ ($\mathbf{X}$ and $\mathbf{Z}_{\mathscr{S}}$, where $\mathscr{S} \subset \{1, \cdots, m\}$ of fixed size $s < m$, and $\mathbf{Z}_{\mathscr{S}}$ is the $n \times s$ matrix of $\mathbf{Z}$-columns indexed by $\mathscr{S}$). Then two *best* subsets of s variables in $\mathbf{Z}$ may be characterized as minimizing

$$\text{(1.1)} \qquad \operatorname{tr} \hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}}_{\mathscr{S}} \quad \text{and} \quad |\hat{\mathbf{\Sigma}}_{\mathscr{S}}|/|\hat{\mathbf{\Sigma}}|$$

over all subsets $\mathscr{S} \subset \{1, \cdots, m\}$ of size s. We show that the significance tests for including-excluding these two best subsets are admissible proper Bayes decision rules for fixed effect variables in $\mathbf{Z}$. The proof of these results follows the approach of Kiefer and Schwartz (1965). If the choice of independent variables is allowed to encompass random and mixed effects in $\mathbf{Z}$, then the latter test is shown to be admissible proper Bayes when restricted to a class of location and scale invariant tests.

In the case of univariate multiple regression, criterion (1.1) has been used almost exclusively to determine the best subset of s independent variables. For other criteria see Allen (1971, 1974), Schmidt (1973), and Stone (1974). Apparently the justification for this use has been based on the fact that if the true model can be selected from the matrix $\mathbf{Z}$, then it is determined by the pair $(\mathscr{S}, s)$ with smallest s for which $E(\hat{\sigma}^2_{\mathscr{S}})$ attains its minimum.

Received April 16, 1979; revised February 2, 1981.

AMS 1970 *subject classifications.* Primary 62C25, 62C10; secondary 62J05, 62G35.

Key words and phrases. Admissibility, Bayesian decision rule, variable selection, outliers, slippage tests, multivariate regression, Models I and II, invariance, Studentized residuals.

Now this use can be justified, at least for the choice of the best $\mathscr{S}$ of fixed size s, on the admissible Bayes character of the related significance test.

Although the amount of computation necessary to minimize (1.1) is much greater than that required by forward selection and backward elimination techniques, the latter have been found in general to be unsound as demonstrated in Lindley (1968) and Hocking (1976). Fortunately, reasonably efficient software has been developed to minimize (1.1) in univariate regression by Beale, Kendall, and Mann (1967), LaMotte and Hocking (1970), and Newton and Spurrel (1967). For additional references and a general survey of variable selection techniques see Hocking (1976).

Special cases of this general selection problem include outlier detection in the normal multivariate regression model and the related problem of slippage detection. By using special independent dummy variables in the variable selection problem, the final sections show the admissibility of *Studentized residual* outlier detection criteria and their related slippage tests. These tests are derived using models that explain outliers and slippage with locational biases, inflated variances, or a combination of the two.

2. Notation. Throughout this paper we will be concerned with the following model or a special case thereof:

$$\mathbf{Y}_{(n\times k)} = \mathbf{X}_{(n\times p)}\cdot\mathbf{\Theta}_{(p\times k)} + \mathbf{Z}_{(n\times m)}\cdot\mathbf{\Delta}_{(m\times k)} + \mathbf{E}_{(n\times k)} \tag{2.1}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are fixed, $\mathbf{\Theta}$ and $\mathbf{\Delta}$ are fixed parameters in Section 3, and either fixed or random normal effects in Section 4, and the rows of $\mathbf{E}$ are independent with joint density

$$f(\mathbf{E}\,|\,\mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-n/2}\text{etr}\,-\tfrac{1}{2}(\mathbf{\Sigma}^{-1}\mathbf{E}'\mathbf{E}) \tag{2.2}$$

where $\text{etr}(\cdot) = \exp\{\text{tr}(\cdot)\}$. Let $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)'$ and $\mathbf{E} = (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n)'$ so that $\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n$ are i.i.d. with law $\text{MVN}_k(\mathbf{0}, \mathbf{\Sigma} > \mathbf{0})$, where k is dimensionality and $\mathbf{\Sigma} > \mathbf{0}$ denotes positive definiteness. Let $\mathbf{\Delta} = (\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m)'$ and $\mathbf{Z} = (\mathbf{z}_1, \cdots, \mathbf{z}_m)$.

We assume the idealized setting where $\mathbf{X}$ is known to be relevant in the regression and all other relevant variables along with others are included in the pool $\mathbf{Z}$. It is also assumed that the choice of variables from $\mathbf{Z}$ is relevant to all k dimensions of $\mathbf{Y}$. This is not a restrictive assumption, however, since uninvolved dimensions may be added to the regression by conditioning as shown in Section 7.

Let $\mathscr{S} = \{i(1), \cdots, i(s)\} \subset \{1, \cdots, m\}$ index a subset of variables of size s, and suppose $\mathbf{Z}_{\mathscr{S}} = (\mathbf{z}_{i(1)}, \cdots, \mathbf{z}_{i(s)})$ is $n \times s$, $\mathbf{\Delta}_{\mathscr{S}} = (\boldsymbol{\delta}_{i(1)}, \cdots, \boldsymbol{\delta}_{i(s)})'$ is $s \times k$, and $\mathbf{X}_{\mathscr{S}} = (\mathbf{X}, \mathbf{Z}_{\mathscr{S}})$ is $n \times (p + s)$ and of full rank $p + s$ for any $\mathscr{S}$. Let

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \qquad \mathbf{M}_{\mathscr{S}} = \mathbf{I} - \mathbf{X}_{\mathscr{S}}(\mathbf{X}_{\mathscr{S}}'\mathbf{X}_{\mathscr{S}})^{-1}\mathbf{X}_{\mathscr{S}}',$$

$$\hat{\mathbf{S}} = \mathbf{Y}'\mathbf{M}\mathbf{Y} = (n-p)\hat{\mathbf{\Sigma}} \qquad \hat{\mathbf{S}}_{\mathscr{S}} = \mathbf{Y}'\mathbf{M}_{\mathscr{S}}\mathbf{Y} = (n-p-s)\hat{\mathbf{\Sigma}}_{\mathscr{S}},$$

$$\hat{\mathbf{\Theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad \begin{pmatrix}\hat{\mathbf{\Theta}}_{\mathscr{S}}\\ \hat{\mathbf{\Delta}}_{\mathscr{S}}\end{pmatrix} = (\mathbf{X}_{\mathscr{S}}'\mathbf{X}_{\mathscr{S}})^{-1}\mathbf{X}_{\mathscr{S}}'\mathbf{Y},$$

and

$$\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{\Theta}} \qquad \hat{\mathbf{E}}_{\mathscr{S}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{\Theta}}_{\mathscr{S}} - \mathbf{Z}_{\mathscr{S}}\hat{\mathbf{\Delta}}_{\mathscr{S}}.$$

Square roots of positive definite matrices will be the unique symmetric positive definite ones. With an abuse of notation, the function $L(\cdot\,|\,\cdot)$ will refer to the likelihood of an arbitrary first argument given parameters in the second argument. Statements which are true with probability one will simply be stated as true so that the final Bayes decision rules are true with probability one.

In the special univariate case,

$$\mathbf{y}_{(n\times 1)} = \mathbf{X}_{(n\times p)}\cdot\boldsymbol{\theta}_{(p\times 1)} + \mathbf{Z}_{(n\times m)}\cdot\boldsymbol{\delta}_{(m\times 1)} + \boldsymbol{\epsilon}_{(n\times 1)},$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)'$ is i.i.d. $\mathscr{N}(0, \sigma^2)$, and

$$\hat{\sigma}^2 = \mathbf{y}'\mathbf{M}\mathbf{y}/(n-p) \text{ and } \hat{\sigma}_{\mathscr{S}}^2 = \mathbf{y}'\mathbf{M}_{\mathscr{S}}\mathbf{y}/(n-p-s).$$

3. Variable Selection in Model I. In the following analysis, assume $\mathbf{\Delta}$ represents some fixed unknown parameter matrix and the "true" model is a special case of (2.1) including either s or 0 of the variables from the pool $\mathbf{Z}$. The multidecision question we wish to consider is the following: Which s variables of $\mathbf{Z}$ (if any) best describe (in terms of an admissible decision rule) the model that has generated our data? An answer to such a question makes no claim to determining s, the number of variables in the true model. However, it does reduce the selection process considerably, and suggests the statistics upon which the choice of s should be based. This decision problem has $\binom{m}{s} + 1$ possible hypothesized states that may be denoted as

$$(3.1) \qquad \begin{aligned} &H_0: \boldsymbol{\Delta} = \mathbf{0} \\ &H_{\mathscr{S}}: \boldsymbol{\delta}_i \neq \mathbf{0} \quad \text{if} \quad i \in \mathscr{S} \\ &\qquad\;\; = \mathbf{0} \quad \text{if} \quad i \notin \mathscr{S} \end{aligned}$$

for all subsets $\mathscr{S}$ of $\{1, \cdots, m\}$ of size s. It is also required that $n \geq p + k + s$.

We shall let $D_{\mathscr{S}}(D_0)$ be the decision that $H_{\mathscr{S}}(H_0)$ is true and assume a zero-one loss function according to whether we hit-miss the correct decision.

THEOREM 3.1. *Let* $P_{\mathscr{S}} = \operatorname{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathscr{S}})$ *and* $\Lambda_{\mathscr{S}} = |\hat{\boldsymbol{\Sigma}}_{\mathscr{S}}| / |\hat{\boldsymbol{\Sigma}}|$ *for any* $\mathscr{S}$ *of size s. Then for model* (2.1) *and any* $c > 0$, *the decision rules*

$$(3.2) \qquad \begin{aligned} &D_0 \quad \text{if} \quad \min_{\mathscr{S}} P_{\mathscr{S}} \geq c \\ &D_{\mathscr{S}^\star} \quad \text{if} \quad P_{\mathscr{S}^\star} = \min_{\mathscr{S}} P_{\mathscr{S}} < c \end{aligned}$$

and

$$(3.3) \qquad \begin{aligned} &D_0 \quad \text{if} \quad \min_{\mathscr{S}} \Lambda_{\mathscr{S}} \geq c \\ &D_{\mathscr{S}^*} \quad \text{if} \quad \Lambda_{\mathscr{S}^*} = \min_{\mathscr{S}} \Lambda_{\mathscr{S}} < c \end{aligned}$$

are admissible proper Bayes decisions. Rule (3.3) *requires the additional assumption that* $n - p - k - 2s \geq 0$. *Both of the rules results from classes of prior distributions with properties* (1) *and* (2).

(1) *Equal prior weight is given to all hypotheses in* $\{H_{\mathscr{S}}: \#\mathscr{S} = s\}$.

(2) *The conditional prior density of* $(\mathbf{Z}'_{\mathscr{S}}\mathbf{M}\mathbf{Z}_{\mathscr{S}})^{1/2}\boldsymbol{\Delta}_{\mathscr{S}} = \boldsymbol{\mu}_{\mathscr{S}}$ *given* $H_{\mathscr{S}}$ *is the same for all* $\mathscr{S}$. [*See* (3.12) *and* (3.16).]

(3) *For* (3.2) *the prior distribution on* $\boldsymbol{\Delta}_{\mathscr{S}}$, $\boldsymbol{\Theta}$, *and* $\boldsymbol{\Sigma}$ *given* $H_{\mathscr{S}}$ *is constrained to* $b^2\boldsymbol{\Sigma} - \boldsymbol{\Delta}'_{\mathscr{S}}\mathbf{Z}'_{\mathscr{S}}\mathbf{M}\mathbf{Z}_{\mathscr{S}}\boldsymbol{\Delta}_{\mathscr{S}} > \mathbf{0}$, *for arbitrary constant* $b > 0$. *This suggests high discriminatory power between* H_0 *and* $H_{\mathscr{S}^\star}$ *near* H_0. [*See* (3.15).]

The best variable subsets, $\mathscr{S}^\star$ *and* $\mathscr{S}^*$, *are the same if either* $s = 1$ *or* $k = 1$. *In the univariate case* $\mathscr{S}^\star = \mathscr{S}^*$ *is the* $\mathscr{S}$ *which minimizes* $\hat{\sigma}^2_{\mathscr{S}}/\hat{\sigma}^2$.

PROOF. With zero-one loss function, the Bayes rule is to decide the hypothesis with maximum posterior probability. Therefore, we decide

$$(3.4) \qquad \begin{aligned} &D_0 \quad \text{if} \quad \Pr(H_0 \mid \mathbf{Y}) > \max_{\mathscr{S}} \Pr(H_{\mathscr{S}} \mid \mathbf{Y}), \\ &D_{\mathscr{S}^*} \quad \text{if} \quad \Pr(H_{\mathscr{S}^*} \mid \mathbf{Y}) = \max_{\mathscr{S}} \Pr(H_{\mathscr{S}} \mid \mathbf{Y}) > \Pr(H_0 \mid \mathbf{Y}). \end{aligned}$$

Such posterior probabilities may be evaluated by integrating the likelihood functions with respect to prior measures on the parameters. Using property (1) for the priors, then $\Pr(H_{\mathscr{S}} \mid \mathbf{Y}) \propto \Phi(\mathbf{Y} \mid H_{\mathscr{S}})$ where

$$(3.5) \qquad \Phi(\mathbf{Y} \mid H_{\mathscr{S}}) = \int L(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\Theta}, \boldsymbol{\Delta}_{\mathscr{S}}, H_{\mathscr{S}})\, d\Pi(\boldsymbol{\Sigma}, \boldsymbol{\Theta}, \boldsymbol{\Delta}_{\mathscr{S}} \mid H_{\mathscr{S}}),$$

and $\Pr(H_0 \mid \mathbf{Y}) \propto p_0 \Phi(\mathbf{Y} \mid H_0)$ where

$$\Phi(\mathbf{Y} \mid H_0) = \int L(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\Theta}, H_0)\, d\Pi(\boldsymbol{\Sigma}, \boldsymbol{\Theta} \mid H_0) \tag{3.6}$$

and $p_0 > 0$ is a constant.

The quantities (3.5) and (3.6) are evaluated using priors on the parameters that were developed by Kiefer and Schwartz (1965) in the context of testing the two decision problem $H_{\mathscr{S}}$ *vs.* H_0. In this paper, they also derive a test statistic for the multidecision problem of classifying a population. We extend these techniques further to include the multidecision model (3.1).

In order to specify $d\Pi(\boldsymbol{\Sigma}, \boldsymbol{\Theta}, \boldsymbol{\Delta}_{\mathscr{S}} \mid H_{\mathscr{S}})$ we first need to transform the data and consider the canonical form of the likelihood under $H_{\mathscr{S}}$. This is done in order to isolate parameter $\boldsymbol{\Delta}_{\mathscr{S}}$ as the mean of a subset of the transformed data (see Anderson (1958), p. 224). Let $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3)$ be an $n \times n$ orthogonal matrix depending on $\mathscr{S}$ such that

$$\mathbf{V}_1 = \mathbf{M}\mathbf{Z}_{\mathscr{S}}(\mathbf{Z}'_{\mathscr{S}}\mathbf{M}\mathbf{Z}_{\mathscr{S}})^{-1/2} \quad \text{is} \quad n \times s$$

and

$$\mathbf{V}_2 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2} \quad \text{is} \quad n \times p.$$

Then transform $\mathbf{U}_i = \mathbf{V}'_i\, \mathbf{Y}$ for $i = 1, 2, 3$ so that

$$\begin{aligned} E(\mathbf{U}_1) &= (\mathbf{Z}'_{\mathscr{S}}\mathbf{M}\mathbf{Z}_{\mathscr{S}})^{1/2}\, \boldsymbol{\Delta}_{\mathscr{S}} = \boldsymbol{\mu}_{\mathscr{S}} && \text{is} \quad s \times k, \\ E(\mathbf{U}_2) &= (\mathbf{X}'\mathbf{X})^{1/2}\boldsymbol{\Theta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_{\mathscr{S}}\boldsymbol{\Delta}_{\mathscr{S}} = \boldsymbol{\nu}_{\mathscr{S}} && \text{is} \quad p \times k, \\ \text{and} \quad E(\mathbf{U}_3) &= \mathbf{0} && \text{is} \quad (n - p - s) \times k. \end{aligned} \tag{3.7}$$

From this canonical reduction, updating formulae for including additional variables in the model are easily derived in the appendix for later use.

In order to compute $\Phi(\mathbf{Y} \mid H_{\mathscr{S}})$ in (3.5) we use the canonical form of $L(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\Theta}, \boldsymbol{\Delta}_{\mathscr{S}}, H_{\mathscr{S}})$ under $H_{\mathscr{S}}$. This is allowable since if $\mathbf{U}' = (\mathbf{U}'_1, \mathbf{U}'_2, \mathbf{U}'_3) = \mathbf{Y}'\mathbf{V}$ then

$$\begin{aligned} L(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\Theta}, \boldsymbol{\Delta}_{\mathscr{S}}, H_{\mathscr{S}}) &\propto |\boldsymbol{\Sigma}|^{-n/2} \operatorname{etr} -\tfrac{1}{2}(\boldsymbol{\Sigma}^{-1}\mathbf{E}'\mathbf{E}) \\ &= |\boldsymbol{\Sigma}|^{-n/2} \operatorname{etr} -\tfrac{1}{2}\{\boldsymbol{\Sigma}^{-1}(\mathbf{U} - E\mathbf{U})'\mathbf{V}'\mathbf{V}(\mathbf{U} - E\mathbf{U})\} \\ &\propto L(\mathbf{U} \mid \boldsymbol{\Sigma}, \boldsymbol{\mu}_{\mathscr{S}}, \boldsymbol{\nu}_{\mathscr{S}}, H_{\mathscr{S}}). \end{aligned}$$

To compute $\Phi(\mathbf{Y} \mid H_0)$ in (3.6) we can use the canonical form of the likelihood for any $\mathscr{S}$.

Now it is possible to specify the priors that result in rules (3.2) and (3.3). For both rules the priors under $H_{\mathscr{S}}$ will be specified on $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_{\mathscr{S}}$, and $\boldsymbol{\nu}_{\mathscr{S}}$ rather than $\boldsymbol{\Sigma}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\Delta}_{\mathscr{S}}$.

Note that $\hat{\mathbf{S}}_{\mathscr{S}} = \mathbf{U}'_3\, \mathbf{U}_3$ and $\hat{\mathbf{S}} = \mathbf{U}'_1\, \mathbf{U}_1 + \mathbf{U}'_3\, \mathbf{U}_3$ (see also the Appendix), so that both decision rules do not depend directly on $\mathbf{U}_2$, and $\mathbf{U}_2$ does not vary with $\mathscr{S}$. This suggests using prior distributions on $\boldsymbol{\Sigma}$ and $\boldsymbol{\nu}_{\mathscr{S}}$ such that the same factor of $\mathbf{U}_2$ enters into $\Phi(\mathbf{Y} \mid H_0)$ and $\Phi(\mathbf{Y} \mid H_{\mathscr{S}})$ for all $\mathscr{S}$. This has the effect of eliminating $\mathbf{U}_2$ and consideration of the nuisance parameter $\boldsymbol{\Theta}$ from the decision problem. Such priors are used for both decision rules and are given in Lemma 3.1 of Kiefer and Schwartz (1965) as follows: Suppose $\boldsymbol{\Sigma}^{-1} = \mathbf{C} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ with prior probability one, where $\mathbf{C} > \mathbf{0}$ is $k \times k$ with some prior distribution and $\boldsymbol{\Lambda}$ is $k \times s$ with prior density to be specified later. Also let $d\Pi(\boldsymbol{\nu}_{\mathscr{S}} \mid \mathbf{C}, \boldsymbol{\Lambda}, \boldsymbol{\Delta}_{\mathscr{S}}, H_{\mathscr{S}})$ be such that with prior probability one

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}'_{\mathscr{S}} = \boldsymbol{\Lambda}\boldsymbol{\Gamma}', \tag{3.8}$$

where $\boldsymbol{\Gamma}$ is $p \times s$ and consists of conditionally i.i.d. row vectors given $\mathbf{C}$, $\boldsymbol{\Lambda}$, and $H_{\mathscr{S}}$ each having law $\mathrm{MVN}_s(\mathbf{0}, \mathbf{B}^{-1})$ where

$$\mathbf{B} = \mathbf{I}_s - \boldsymbol{\Lambda}'(\mathbf{C} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}')^{-1}\boldsymbol{\Lambda}. \tag{3.9}$$

The independence of the columns of Γ' results in the conditional independence of the columns of $\nu'_{\mathscr{S}}$ given $\mathbf{C}$, Λ, and $H_{\mathscr{S}}$.

This collection of conditional priors has the effect of giving Θ a different distribution under each $H_{\mathscr{S}}$ even though $H_{\mathscr{S}}$ makes no reference to Θ. While such a property is hardly desirable, the choice of priors by Kiefer and Schwartz was not intended to express prior beliefs. The Bayesian framework is simply used to demonstrate the admissibility of (3.2) and (3.3).

Under the canonical reduction above for any $\mathscr{S}$, $E(\mathbf{U}_2 \mid H_0) = (\mathbf{X}'\mathbf{X})^{1/2}\boldsymbol{\theta} = \boldsymbol{\nu}_0$. Under H_0 and given $\mathbf{C}$ and Λ, we let $d\Pi(\boldsymbol{\nu}_0 \mid \mathbf{C}, \Lambda, H_0) = d\Pi(\boldsymbol{\nu}_{\mathscr{S}} \mid \mathbf{C}, \Lambda, H_{\mathscr{S}})$.

The effect of such priors on the calculations in (3.5) and (3.6) may be viewed through their effects on $L(\mathbf{U}_2 \mid \boldsymbol{\nu}_{\mathscr{S}}, H_{\mathscr{S}})$ integrated over Γ. Since

$$\int L(\mathbf{U}_2 \mid \boldsymbol{\nu}_{\mathscr{S}}, \Lambda, \mathbf{C}, H_{\mathscr{S}})\, d\Pi(\Gamma \mid \mathbf{C} \mid, \Lambda, H_{\mathscr{S}}) \propto |\mathbf{C}|^{p/2} \operatorname{etr}\{-\tfrac{1}{2}(\mathbf{C}\mathbf{U}_2'\mathbf{U}_2)\}, \tag{3.10}$$

then this is the only factor contributing to $\Pr\{H_{\mathscr{S}} \mid \mathbf{Y}\}$ from $\mathbf{U}_2$. Factor (3.10) does not depend on $\mathscr{S}$ and appears in the computation of $\Pr(H_{\mathscr{S}} \mid \mathbf{Y})$ for any $\mathscr{S}$ and $\Pr(H_0 \mid \mathbf{Y})$ so that $\mathbf{U}_2$ and $\boldsymbol{\nu}_{\mathscr{S}}$ are effectively removed from consideration.

To show that criterion (3.2) is Bayes, consider the following additional priors on $\boldsymbol{\mu}_{\mathscr{S}}$ and Σ taken from equation (4.1) of Kiefer and Schwartz (1965). Under $H_{\mathscr{S}}$ and H_0 let

$$\Sigma^{-1} = \mathbf{C} + \Lambda\Lambda' \tag{3.11}$$

and under $H_{\mathscr{S}}$ let

$$\boldsymbol{\mu}_{\mathscr{S}} = (\mathbf{Z}_{\mathscr{S}}'\mathbf{M}\mathbf{Z}_{\mathscr{S}})^{1/2}\Delta_{\mathscr{S}} = b\Lambda'\Sigma \tag{3.12}$$

for random $k \times s$ matrix Λ and arbitrary constant $b > 0$. Now, under the various hypotheses, let the priors on Λ given $\mathbf{C}$ be absolutely continuous on ks-dimensional space with densities

$$d\Pi(\Lambda \mid \mathbf{C}, H_{\mathscr{S}}) \propto |\mathbf{C}|^{(n-p-s)/2} |\mathbf{C} + \Lambda\Lambda'|^{-(n-p)/2} \operatorname{etr}[b^2\{\Lambda'(\mathbf{C} + \Lambda\Lambda')\}/2]\, d\Lambda \tag{3.13}$$

and

$$d\Pi(\Lambda \mid \mathbf{C}, H_0) \propto |\mathbf{C}|^{(n-p-s)/2} |\mathbf{C} + \Lambda\Lambda'|^{-(n-p)/2}\, d\Lambda. \tag{3.14}$$

These densities are integrable when $n - p - k - s \geq 0$. Let $\mathbf{C}$ have a Wishart $(\mathbf{I}_k, k + s)$ law under all hypotheses. Excluding the contribution from $\mathbf{U}_2$ and after integrating over Λ and $\mathbf{C}$ then

$$\Pr\{H_{\mathscr{S}} \mid \mathbf{Y}\} \propto |\hat{\mathbf{S}}|^{-s/2} \operatorname{etr}\{-\tfrac{1}{2} b^2 (\hat{\mathbf{S}}^{-1}\hat{\mathbf{S}}_{\mathscr{S}})\}$$

and

$$\Pr(H_0 \mid \mathbf{Y}) \propto p_0 |\hat{\mathbf{S}}|^{-s/2}$$

so that (3.2) is a Bayes rule.

The distribution on $\mathbf{C} > \mathbf{0}$ is required in order that the prior support on $\Delta_{\mathscr{S}}$ and Σ under $H_{\mathscr{S}}$ is not degenerate. These priors are supported on

$$\Sigma^{-1} - \Lambda\Lambda' = \Sigma^{-1} - \Sigma^{-1}\Delta_{\mathscr{S}}'\mathbf{Z}_{\mathscr{S}}'\mathbf{M}\mathbf{Z}_{\mathscr{S}}\Delta_{\mathscr{S}}\Sigma^{-1}/b^2 > \mathbf{0}, \tag{3.15}$$

or when $b^2\Sigma - \Delta_{\mathscr{S}}'\mathbf{Z}_{\mathscr{S}}'\mathbf{M}\mathbf{Z}_{\mathscr{S}}\Delta_{\mathscr{S}} > \mathbf{0}$. Note from (A2) of the Appendix that the covariance of the ith and jth columns of $\hat{\Delta}_{\mathscr{S}}$ is $(\mathbf{Z}_{\mathscr{S}}'\mathbf{M}\mathbf{Z}_{\mathscr{S}})^{-1}\sigma_{ij}$, where $(\sigma_{ij}) = \Sigma$ so that the support of $\Delta_{\mathscr{S}}$ and Σ is a *Studentized region less than* $b^2\mathbf{I}_k$.

We now show that rule (3.3) is proper Bayes when $n - p - k - 2s \geq 0$. Priors $d\Pi(\boldsymbol{\nu}_{\mathscr{S}} \mid \mathbf{C}, \Lambda, \Delta_{\mathscr{S}}, H_{\mathscr{S}})$ are as before in (3.8). The priors on $\Delta_{\mathscr{S}}$ and Σ given $H_{\mathscr{S}}$ are taken from expression (4.4) in Kiefer and Schwartz. Let $\Sigma^{-1} = \mathbf{C} + \Lambda\Lambda'$ where $d\Pi(\mathbf{C} \mid H_{\mathscr{S}}) = d\Pi(\mathbf{C} \mid H_0)$ is the Wishart $(\mathbf{I}_k, n + s)$ density for any $\mathscr{S}$. Let $d\Pi(\boldsymbol{\mu}_{\mathscr{S}} \mid \mathbf{C}, \Lambda, H_{\mathscr{S}})$ be such that

$$\boldsymbol{\mu}_{\mathscr{S}} = \mathbf{D}\Lambda'\Sigma \tag{3.16}$$

where $\mathbf{D}$ is $s \times s$ and consists of i.i.d row vectors each of law MVN($\mathbf{0}$, $\mathbf{B}^{-1}$), where $\mathbf{B}$ is given in (3.9). Let

$$d\Pi(\Lambda \mid \mathbf{C}, H_{\mathscr{S}}) \propto |\mathbf{C}|^{(n-p-s-k)/2} |\mathbf{C} + \Lambda\Lambda'|^{-(n-p-s)/2}\, d\Lambda \tag{3.17}$$

and suppose $d\Pi(\Lambda \mid \mathbf{C}, H_0)$ is given in (3.14). The requirement that $n - p - k - 2s \geq 0$ is necessary for the integrability of (3.17). Then we may compute (3.5) and (3.6) by completing the square on $\mathbf{D}$, and integrating successively $d\mathbf{D}$, $d\Lambda$, and $d\mathbf{C}$ so that

$$\Pr(H_{\mathscr{S}} \mid \mathbf{Y}) \propto |\hat{\mathbf{S}}_{\mathscr{S}}|^{-s/2} \quad \text{and} \quad \Pr(H_0 \mid \mathbf{Y}) \propto p_0 |\hat{\mathbf{S}}|^{-s/2} \tag{3.18}$$

and (3.3) is a Bayes rule. □

In contrast to decision rule (3.2), the support of the prior distributions leading to rule (3.3) includes the entire parameter space.

4. Variable Selection in Model II and Mixed Models.

Now let the parameter $\boldsymbol{\Delta}$ be a random variable that allows for normally distributed random effects in the design. The model we assume here allows the parameters in $\boldsymbol{\theta}$ and $\boldsymbol{\Delta}$ to be either fixed or random effects. This was not possible in Section 3 because of the type of prior distributions assigned to the parameters under each hypothesis. We again consider the selection of s (if any) independent variables from the pool $\mathbf{Z}$ in the hypothesis framework of (3.1), where now each $\boldsymbol{\delta}_i = \mathbf{0}$ must be qualified as true with probability one.

For this model the admissibility of selection criterion (3.3) is again demonstrated using Bayesian techniques. This admissibility property is restricted to hold over the class of decision rules that are invariant to the following transformations:

1. Translation of $\mathbf{Y}$ by a matrix of column vectors, each in the column space of $\mathbf{X}$. This allows invariance to the rescaling of independent variables known to be contained in the true model.

2. Non-singular $k \times k$ transformations of the rows of $\mathbf{Y}$ which includes rescaling of the data $\mathbf{Y}$.

Rule (3.3) results from random effects and priors with these properties:

1. Equal prior weights are given to all hypotheses in $\{H_{\mathscr{S}}\colon \#\mathscr{S} = s\}$.

2. If $\boldsymbol{\Delta}_{\mathscr{S}} = (\boldsymbol{\delta}_{i(1)}, \cdots, \boldsymbol{\delta}_{i(s)})'$ and $\boldsymbol{\delta}'_{\mathscr{S}} = (\boldsymbol{\delta}'_{i(1)}, \cdots, \boldsymbol{\delta}'_{i(s)})$ is $1 \times ks$, then the distribution of $\boldsymbol{\delta}_{\mathscr{S}}$, given $\boldsymbol{\Sigma} > \mathbf{0}$ and $\mathbf{Q} > \mathbf{0}$ which is $s \times s$, is $\mathrm{MVN}_{ks}(\mathbf{0}, \mathbf{Q} \otimes \boldsymbol{\Sigma})$.

3. The variance-covariance scalars of $\mathbf{Q}$ are assigned conditional priors such that the distribution of $\mathbf{Q}(\mathbf{Z}'_{\mathscr{S}}\mathbf{M}\mathbf{Z}_{\mathscr{S}})$ under $H_{\mathscr{S}}$ is the same for all $\mathscr{S}$. Such priors make $E(\hat{\boldsymbol{\Sigma}} \mid \boldsymbol{\Sigma}, H_{\mathscr{S}})$ independent of $\mathscr{S}$; thus $\hat{\boldsymbol{\Sigma}}$ is made a useful benchmark for comparing the various $\hat{\boldsymbol{\Sigma}}_{\mathscr{S}}$ over all $\mathscr{S}$. This result follows by noting that

$$\begin{aligned} E(\hat{\boldsymbol{\Sigma}} \mid \boldsymbol{\Sigma}, \boldsymbol{\Delta}_{\mathscr{S}}, H_{\mathscr{S}}) &= \boldsymbol{\Sigma} + \boldsymbol{\Delta}'_{\mathscr{S}}\mathbf{Z}'_{\mathscr{S}}\mathbf{M}\mathbf{Z}_{\mathscr{S}}\boldsymbol{\Delta}_{\mathscr{S}}/(n-p) \quad \text{so that} \\ E(\hat{\boldsymbol{\Sigma}} \mid \boldsymbol{\Sigma}, \mathbf{Q}, H_{\mathscr{S}}) &= \{1 + \operatorname{tr} \mathbf{Q}\mathbf{Z}'_{\mathscr{S}}\mathbf{M}\mathbf{Z}_{\mathscr{S}}/(n-p)\}\boldsymbol{\Sigma} \end{aligned} \tag{4.1}$$

and therefore the prior mean of $\hat{\boldsymbol{\Sigma}}$ is the same under each $H_{\mathscr{S}}$.

THEOREM 4.1. *Let $\Lambda_{\mathscr{S}} = |\hat{\boldsymbol{\Sigma}}_{\mathscr{S}}|/|\hat{\boldsymbol{\Sigma}}|$. Then, under model* (2.1), *hypothesis framework* (3.1), *and mixed model assumptions on $\boldsymbol{\Theta}$ and $\boldsymbol{\Delta}$, the rule*

$$\begin{aligned} &\mathrm{D}_0 && \text{if } \min_{\mathscr{S}} \Lambda_{\mathscr{S}} \geq c \\ &\mathrm{D}_{\mathscr{S}^*} && \text{if } \Lambda_{\mathscr{S}^*} = \min_{\mathscr{S}} \Lambda_{\mathscr{S}} < c \end{aligned} \tag{4.2}$$

is admissible Bayes in the class of invariant decisions characterized above.

PROOF. Suppose $\boldsymbol{\Gamma}_{\mathscr{S}} = \boldsymbol{\Delta}_{\mathscr{S}}\boldsymbol{\Sigma}^{-1/2} = (\mathbf{g}_1, \cdots, \mathbf{g}_k)$ is $s \times k$ and $\mathbf{g}'_{\mathscr{S}} = (\mathbf{g}'_1, \cdots, \mathbf{g}'_k)$ is $1 \times ks$.

Then the prior distribution of $\mathbf{\Delta}_{\mathscr{S}}$ given $H_{\mathscr{S}}$ is such that $\{\mathbf{g}_1, \cdots, \mathbf{g}_k\}$ are i.i.d. with $\text{MVN}_s(\mathbf{0}, \mathbf{Q})$ law. This distribution will be used below.

If a decision rule is to be invariant to the transformations above, then it must be a function of the maximal invariant. In our invariant Bayes approach, this justifies the use of the likelihood for the maximal invariant rather than $\mathbf{Y}$. The use of the maximal invariant likelihood function also results from the assumption of Jeffreys' (1937) invariant improper prior $d\Pi(\mathbf{\Sigma}, \mathbf{\Theta} \mid H_{\mathscr{S}}) = d\Pi(\mathbf{\Sigma}, \mathbf{\Theta} \mid H_0) \propto |\mathbf{\Sigma}|^{(k+1)/2}$. We denote the maximal invariant as $\mathbf{v}' = (\mathbf{v}_1', \cdots, \mathbf{v}_{n-p-k}')$, which is $1 \times (n-p-k)k$, and derive it below.

The matrix $\hat{\mathbf{E}} = \mathbf{MY}$ is invariant to the type 1 transformations given above. Let $\mathbf{\Omega}$ be an $n \times n$ orthogonal matrix such that

$$\mathbf{\Omega}'\mathbf{M}\mathbf{\Omega} = \begin{pmatrix} \mathbf{I}_{n-p} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Then we may remove the singularities in $\hat{\mathbf{E}}$ by transforming so that $\mathbf{\Omega}'\hat{\mathbf{E}} = (\mathbf{w}_1, \cdots, \mathbf{w}_{n-p}, \mathbf{0}, \cdots, \mathbf{0})'$. Let $\mathbf{W} = (\mathbf{w}_{n-p-k+1}, \cdots, \mathbf{w}_{n-p})$ be $k \times k$ and non-singular with probability one. Transform so that

$$\mathbf{v}_i = \mathbf{W}^{-1}\mathbf{w}_i, \quad i = 1, \cdots, n-p \tag{4.3}$$

and $(\mathbf{v}_{n-p-k+1}, \cdots, \mathbf{v}_{n-p}) = \mathbf{I}_k$. The $\mathbf{v}_i$'s are invariant to type 2 transformations and it is easy to show that $\mathbf{v}' = (\mathbf{v}_1', \cdots, \mathbf{v}_{n-p-k}')$ is the maximal invariant.

Straightforward Jacobian transformations can now be used to show that

$$L(\mathbf{v} \mid \mathbf{\Delta}_{\mathscr{S}}, H_{\mathscr{S}}) \propto \int \|\mathbf{T}\|^{n-p-k} \exp[-\tfrac{1}{2}\{\textstyle\sum_{i=1}^{n-p} (\mathbf{T}'\mathbf{v}_i - \boldsymbol{\eta}_i)'(\mathbf{T}'\mathbf{v}_i - \boldsymbol{\eta}_i)\}]\, d\mathbf{T}, \tag{4.4}$$

where $\mathbf{T} = \mathbf{W}'\mathbf{\Sigma}^{-1/2}$, $\boldsymbol{\eta}_i' = E(\mathbf{w}_i' \mid \mathbf{\Delta}_{\mathscr{S}}, H_{\mathscr{S}})\mathbf{\Sigma}^{-1/2} = i^{\text{th}}$ row of $\mathbf{\Omega}'\mathbf{M}\mathbf{Z}_{\mathscr{S}}\mathbf{\Gamma}_{\mathscr{S}}$, and $\|\mathbf{T}\| = |\det \mathbf{T}|$.

Quite a lot of tedious matrix algebra and transformations are required to arrive at $\Pr(H_{\mathscr{S}} \mid \mathbf{v}, \mathbf{Q})$ from (4.4). It is expedient to simply describe the computations. Before integrating $d\mathbf{T}$, express the exponential of (4.4) as a quadratic form in $\mathbf{g}_{\mathscr{S}}$. Since the exponential of $d\Pi(\mathbf{\Delta}_{\mathscr{S}} \mid \mathbf{Q}, H_{\mathscr{S}})$ or $d\Pi(\mathbf{g}_{\mathscr{S}} \mid \mathbf{Q}, H_{\mathscr{S}})$ is also a quadratic form in $\mathbf{g}_{\mathscr{S}}$, then the integrated likelihood over $d\mathbf{g}_{\mathscr{S}}$ is a normal integral which reduces to

$$\Pr(H_{\mathscr{S}} \mid \mathbf{v}, \mathbf{Q}) \propto |\mathbf{Q}\mathbf{D}_{\mathscr{S}}|^{-k/2} \int \|\mathbf{T}\|^{n-p-k} \operatorname{etr}[-\tfrac{1}{2}\{\mathbf{T}\mathbf{T}'\mathbf{W}^{-1}\hat{\mathbf{E}}'(\mathbf{I}_k - \mathbf{Z}_{\mathscr{S}}\mathbf{D}_{\mathscr{S}}^{-1}\mathbf{Z}_{\mathscr{S}}')\hat{\mathbf{E}}\mathbf{W}^{-1}\}]\, d\mathbf{T} \tag{4.5}$$

where $\mathbf{D}_{\mathscr{S}} = \mathbf{Z}_{\mathscr{S}}\mathbf{M}\mathbf{Z}_{\mathscr{S}} + \mathbf{Q}^{-1}$. Now let $\mathbf{G} = \mathbf{T}\mathbf{T}'$, $\mathbf{H} = h(\mathbf{T})$ be a $1-1$ transformation, where h is a differentiable transformation mapping k^2-space onto $k(k-1)/2$-space. Then by Lemma 13.3.1 of Anderson (1958), p. 319,

$$\int \|\partial\mathbf{T}/\partial(\mathbf{G}, \mathbf{H})\|\, d\mathbf{H} \propto |\mathbf{G}|^{-1/2}.$$

Using this, along with the fact that (4.5) is a function of $\mathbf{G}$ only and not $\mathbf{H}$, then (4.5) takes the form of a Wishart integral over $d\mathbf{G}$ which integrates to

$$\Pr(H_{\mathscr{S}} \mid \mathbf{v}, \mathbf{Q}) \propto |\mathbf{Q}\mathbf{D}_{\mathscr{S}}|^{-k/2} |\mathbf{W}^{-2}\hat{\mathbf{S}}|^{-(n-p)/2} |\mathbf{I}_k - \hat{\mathbf{S}}^{-1}\hat{\mathbf{E}}'\mathbf{Z}_{\mathscr{S}}\mathbf{D}_{\mathscr{S}}^{-1}\mathbf{Z}_{\mathscr{S}}'\hat{\mathbf{E}}|^{-(n-p)/2}. \tag{4.6}$$

The posterior probability of $\mathbf{H}_0$ can be derived as

$$\Pr(\mathbf{H}_0 \mid \mathbf{v}) \propto p_0 |\mathbf{W}^{-2}\hat{\mathbf{S}}|^{-(n-p)/2}. \tag{4.7}$$

What remains is to integrate (4.6) over $d\Pi p(\mathbf{Q} \mid H_{\mathscr{S}})$. We assign a prior distribution to $\mathbf{Q}$ under $H_{\mathscr{S}}$ as follows. Let $(\mathbf{Z}_{\mathscr{S}}'\mathbf{M}\mathbf{Z}_{\mathscr{S}})^{-1}\mathbf{Q}^{-1} = \mathbf{N}\mathbf{N}'$ so that $\mathbf{N}$ is an $s \times s$ square root with prior density

$$d\Pi(\mathbf{N} \mid H_{\mathscr{S}}) \propto |\mathbf{I}_s + \mathbf{N}\mathbf{N}'|^{-(n-p-k)/2} d\mathbf{N} \tag{4.8}$$

that is integrable when $n - p - k \geq 2s$. After some matrix algebra which involves using (A.3) of the Appendix, then $\int \Pr(H_{\mathscr{S}} | \mathbf{v}, \mathbf{N}) d\Pi(\mathbf{N} | H_{\mathscr{S}})$ takes the form of a multivariate beta integral which integrates to

$$\Pr(H_{\mathscr{S}} | \mathbf{v}) \propto | \mathbf{W}^{-2}\hat{\mathbf{S}} |^{-(n-p)/2} \{ | \hat{\mathbf{S}}_{\mathscr{S}} | / | \hat{\mathbf{S}} | \}^{-(n-p-k-s)/2} \tag{4.9}$$

so (4.2) is an invariant Bayes decision rule. □

Up to this point we have ignored $\mathbf{\Sigma} > \mathbf{0}$ because of our consideration of invariant decision rules. Suppose prior information about $\mathbf{\Sigma}$ exists and we wish to include this additional information in the decision rule and maintain our invariant approach. This may be accomplished by assuming our prior information is based on *hard data*, $\hat{\mathbf{S}}_0 = \sum_{i=1}^{\nu} \mathbf{w}_i^0 \mathbf{w}_i^{0\prime}$, having a Wishart $(\mathbf{\Sigma}, \nu)$ distribution that results from the i.i.d. sequence $\{\mathbf{w}_1^0, \cdots, \mathbf{w}_\nu^0\}$ having a $\mathscr{N}(\mathbf{0}, \mathbf{\Sigma})$ law. Such information may be incorporated into the likelihood, by replacing the transformation (4.3) with

$$\mathbf{v}_i = \mathbf{W}^{-1}\mathbf{w}_i \qquad i = 1, \cdots, n-p$$

$$\mathbf{v}_j^0 = \mathbf{W}^{-1}\mathbf{w}_j^0, \qquad j = 1, \cdots, \nu.$$

The net effect of this replacement is the substitution of $\hat{\mathbf{S}} + \hat{\mathbf{S}}_0$ for $\hat{\mathbf{S}}$ and $n + \nu$ for n in the analysis.

COROLLARY 4.2. *If, under the conditions of Theorem* 4.1, *additional prior information regarding* $\mathbf{\Sigma}$ *is available in the form* $\hat{\mathbf{S}}_0 \sim$ *Wishart* $(\mathbf{\Sigma}, \nu)$ *independently of* $\mathbf{Y}$, *then the decision rule* (4.2) *with*

$$\Lambda_{\mathscr{S}} = | (\hat{\mathbf{S}}_{\mathscr{S}} + \hat{\mathbf{S}}_0)/(n + \nu - p - s) | / | (\hat{\mathbf{S}} + \hat{\mathbf{S}}_0)/(n + \nu - p) |$$

is an invariant admissible Bayes rule.

While s has been fixed throughout this discussion, it may also be assigned a prior distribution. Suppose s has prior support $\{0, 1, \cdots, s_0\}$ and for fixed s make all the assumptions that were necessary to obtain rule (4.2). Assume an hypothesis space which includes $\cup_{s=0}^{s_0} \{H_{\mathscr{S}}: \#\mathscr{S} = s\}$ and a zero-one loss function. Then for any function $\kappa(s)$, the rule which determines $\mathscr{S}$ and s by maximizing

$$\kappa(s) + (n - p - k - s)\ln[| \hat{\mathbf{S}} | / | \hat{\mathbf{S}}_{\mathscr{S}} |]$$

over all $\{\mathscr{S}: \#\mathscr{S} = s\}$ and $s \in \{0, 1, \cdots, s_0\}$ is an invariant admissible Bayes rule. A suitable $\kappa(\cdot)$ has not been investigated.

5. Outlier Detection. This section demonstrates the optimality properties of various outlier detection criteria that are based upon the Studentized residuals. These properties will follow from the two previous sections when outlier detection is shown to be equivalent to the selection of special independent dummy variables. This fact has also been used by Dempster (1969) for the computational elimination of individual data points.

We consider model (2.1) and let H_0: $\mathbf{\Delta} = \mathbf{0}$ represent the situation where no outliers are present in the model. Let s be the number of suspected mavericks, and suppose $n - p - k - s \geq 0$ and all $(n - s) \times p$ submatrices of $\mathbf{X}$ are of full rank p. By taking $m = n$ and $\mathbf{Z} = \mathbf{I}_n$ in (2.1), then $H_{\mathscr{S}}$ represents the situation where each data point in $\{ \mathbf{y}_i: i \in \mathscr{S}\}$ has a distinct nonzero bias term added into the model, so that $\mathscr{S} \subset \{1, \cdots, n\}$ indexes a subset of outliers. Under $H_{\mathscr{S}}$, the addition of $\boldsymbol{\delta}_{\mathscr{S}}$ effectively eliminates $\{ \mathbf{y}_i: i \in \mathscr{S}\}$ from the estimator $\hat{\mathbf{\Sigma}}_{\mathscr{S}}$ so that $\hat{\mathbf{\Sigma}}_{\mathscr{S}} = \hat{\mathbf{\Sigma}}(/\mathscr{S})$ is the usual unbiased estimator of $\mathbf{\Sigma}$ fitting the independent variables of $\mathbf{X}$ without using $\{ \mathbf{y}_i: i \in \mathscr{S}\}$. The results of Theorem 3.1 yield the following.

COROLLARY 5.1. *Suppose data is derived from model* (2.1) *as above and either there are s spurious observations caused by biases in location or none at all. Then decision* *rules* (3.2) *and* (3.3) *with*

$$P_{\mathscr{S}} = \operatorname{tr} \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}(/\mathscr{S}) \quad \textit{and} \quad \Lambda_{\mathscr{S}} = |\hat{\boldsymbol{\Sigma}}(/\mathscr{S})| / |\hat{\boldsymbol{\Sigma}}|$$

are admissible proper Bayes outlier detection rules. Both rules result from a class of priors with these properties:

(1) *Equal weights are given to all hypotheses in* $\{H_{\mathscr{S}}: \#\mathscr{S} = s\}$.

(2) *The prior density of* $\mathbf{M}(\mathscr{S})^{1/2}\boldsymbol{\Delta}_{\mathscr{S}}$ *given* $H_{\mathscr{S}}$ *is the same for all* $\mathscr{S}$, *where* $\mathbf{M}(\mathscr{S}) = (\mathbf{M}_{ij}: i, j \in \mathscr{S})$.

(3) *For the test based on* $\{P_{\mathscr{S}}\}$, *the prior distribution on* $\boldsymbol{\Delta}_{\mathscr{S}}$, $\boldsymbol{\Theta}$, *and* $\boldsymbol{\Sigma}$ *given* $H_{\mathscr{S}}$ *has support such that* $b^2\boldsymbol{\Sigma} - \boldsymbol{\Delta}_{\mathscr{S}}'\mathbf{M}(\mathscr{S})\boldsymbol{\Delta}_{\mathscr{S}} > \mathbf{0}$ *for arbitrary* $b > 0$.

The outlier model for this corollary uses a separate independent dummy variable for each possible outlier present and therefore allows for the possibility that there are unrelated causes for multiple outliers. This model was first suggested by Dixon (1953) as Model A.

Special cases of these detection criteria have been proposed by various authors. The list below identifies the author with the model and form of $\Lambda_{\mathscr{S}}$ that was considered.

TABLE 5.1

Author	Normal Model	Criterion
Wilks (1963)	i.i.d. multivariate any k and s	$\lvert\hat{\boldsymbol{\Sigma}}(/\mathscr{S})\rvert / \lvert\hat{\boldsymbol{\Sigma}}\rvert$
Grubbs (1950)	i.i.d. univariate $k = 1$, any s	$\hat{\sigma}^2(/\mathscr{S})/\hat{\sigma}^2$
Thompson (1935)	i.i.d. univariate $k = 1, s = 1$	$(y_i - \bar{y})^2/\hat{\sigma}^2$
Gentleman and Wilk (1975)	2-factor ANOVA $k = 1$, any s	$\hat{\sigma}^2(/\mathscr{S})/\hat{\sigma}^2$
Srikantan (1961); Beckman and Trussell (1974)	univariate regression $k = 1, s = 1$	$(y_i - x_i'\hat{\boldsymbol{\theta}})^2/\{(1 - A_{ii})\hat{\sigma}^2\}$

Here $A_{ii} = i^{\text{th}}$ diagonal element of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

In the detection of a single outlier, $s = 1$, various optimality properties related to Corollary 5.1 have been shown. Uniform invariant admissibility of the tests above has been shown in the case of Wilk's model ($s = 1$) by Karlin and Truax (1960), for Thompson's model by Paulson (1952) and Kudo (1956a), and for Srikantan's model by Ferguson (1960).

Suppose now that mavericks originate from some common source and it is reasonable to assume the same bias for each outlier. Then a single constant parameter may be used to explain all s possible outliers in the model (2.1). This can be carried out with dummy variables by letting $\mathbf{Z}$ be $n \times \binom{n}{s}$ and consist of columns that include all permutations of s ones and $n - s$ zeros. Then, we are interested in selecting only one variable from pool $\mathbf{Z}$ so that $s = 1$.

COROLLARY 5.2. *Suppose data results from the model* (2.1) *with* $\mathbf{X} = (1, \cdots, 1)'$ *an* $n \times 1$ *location vector. Let there be s spurious observations of equal bias or none at all. Then rules* (3.2) *and* (3.3) *are identical,* $\Lambda_{\mathscr{S}}$ *is a strictly decreasing function of*

$$(\bar{\mathbf{y}}_{\mathscr{S}} - \bar{\mathbf{y}})'\hat{\mathbf{\Sigma}}^{-1}(\bar{\mathbf{y}}_{\mathscr{S}} - \bar{\mathbf{y}}) \tag{5.1}$$

where $\bar{\mathbf{y}}_{\mathscr{S}} = \sum_{i\in\mathscr{S}} \mathbf{y}_i/s$ *and this test is admissible Bayes.*

The univariate case of (5.1) is one of the detection criteria suggested by Thompson (1935). Its invariant admissibility was shown by Murphy (1951) for a one-sided significance test.

Suppose now we assume that mavericks are caused by an inflated error variance as suggested by Dixon's (1953) Model B. Then with $m = n$ and $\mathbf{Z} = \mathbf{I}_n$, the results of Theorem 4.1 apply.

COROLLARY 5.3. *Under model* (2.1) *and the Model B assumption that inflated variances cause mavericks, the rule* (4.2) *with* $\Lambda_{\mathscr{S}} = |\hat{\mathbf{\Sigma}}(/\mathscr{S})|/|\hat{\mathbf{\Sigma}}|$ *is an invariant admissible Bayes rule in the class of decision rules defined in the last section.*

This decision rule results from the random effect $\boldsymbol{\delta}_{\mathscr{S}}$ having a $MVN(\mathbf{0}, \mathbf{Q} \otimes \mathbf{\Sigma})$ law which inflates the covariance of $\mathbf{y}_{\mathscr{S}}$ from $\mathbf{I}s \otimes \Sigma$ to $(\mathbf{I}s + \mathbf{Q}) \otimes \mathbf{\Sigma}$. In addition, the variance-covariance scalars of $\mathbf{Q}$ are assigned priors such that the distribution of $\mathbf{QM}(\mathscr{S})$ under $H_{\mathscr{S}}$ is the same for all $\mathscr{S}$.

A related optimality result has been shown under Model B for $s = 1$. Ferguson (1960) has shown that in the case of an i.i.d. multivariate sample, the test for a single outlier is uniformly invariant admissible.

It is interesting to note the relationship of our Model B development in Section 4 with the Bayesian outlier procedures of Box and Tiao (1968) for $k = 1$. They assume the model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$ and that each ϵ_i can be drawn from one of two sources—a central model $\mathscr{N}(0, \sigma^2)$ and an alternative model $\mathscr{N}(0, r^2\sigma^2)$, $r > 1$ and fixed, with probabilities $(1 - \alpha)$ and α respectively. Using such a model with Jeffreys' priors on $\boldsymbol{\theta}$ and σ, yields an expression for $\Pr\{H_{\mathscr{S}} | \mathrm{v}, \mathbf{Q} = (r^2 - 1)\mathbf{I}s\}$ in (3.15) of Box and Tiao that agrees with our (4.6) taken with $k = 1$, $\mathbf{Q} = (r^2 - 1)\mathbf{I}s$, prior weight $\Pr(H_{\mathscr{S}}) \propto \alpha^s(1 - \alpha)^{n-s}$, and $\mathbf{Z}'_{\mathscr{S}}\mathbf{M}\mathbf{Z}_{\mathscr{S}} = \mathbf{M}(\mathscr{S})$.

When additional information about $\mathbf{\Sigma}$ is available in a *hard data* form, then Corollary 4.2 may be applied. Under the model of Thompson for which $s = 1$, then this result demonstrates the invariant admissibility of the criteria of Quesenberry and David (1961).

If mavericks are caused by both a bias in the location and an inflation of variance then the invariant admissibility of Corollary 5.3 still holds. This is because the likelihood component responsible for the variance inflation and the prior on the bias are independent and in the Bayesian development become unidentifiable. Under $H_{\mathscr{S}}$, if $\boldsymbol{\delta}_{\mathscr{S}}$ is the bias with prior $\mathrm{MVN}(\mathbf{0}, \mathbf{P} \otimes \mathbf{\Sigma})$ and $\boldsymbol{\xi}_{\mathscr{S}}$ is the likelihood component responsible for the inflation so that $\boldsymbol{\xi}_{\mathscr{S}}$ has law $\mathrm{MVN}(\mathbf{0}, \mathbf{Q} \otimes \mathbf{\Sigma})$, then their independence and unidentifiability allow the assumption of random effect $\boldsymbol{\delta}_{\mathscr{S}} + \boldsymbol{\xi}_{\mathscr{S}} \sim \mathrm{MVN}(\mathbf{0}, (\mathbf{P} + \mathbf{Q}) \otimes \mathbf{\Sigma})$. Priors like (4.8) are assigned to $\mathbf{P} + \mathbf{Q}$ and the optimality follows. Kudo (1956b) has shown the invariant admissibility of Thompson's criterion when mavericks result from a bias and a decrease in variance.

Even though the priors of Corollaries 5.1 and 5.3 are design dependent, their use may not be so unreasonable. In these two results $\mathbf{M}(\mathscr{S})^{1/2}\mathbf{\Delta}_{\mathscr{S}}$ and $\mathbf{QM}(\mathscr{S})$ given $H_{\mathscr{S}}$ both have prior distributions which do not change with $\mathscr{S}$. Since the diagonal elements of $\mathbf{M}(\mathscr{S})$ are larger for the more *internal* values of the independent variables, then these priors shift the support of $\mathbf{\Delta}_{\mathscr{S}}$ and $\mathbf{Q}$ closer to the origin for the more internally located subsets $\mathscr{S}$. A priori then, we are more likely to decide $D_{\mathscr{S}}$ if $\mathscr{S}$ is more externally located in terms of the design. The desirability of such a property may be inferred from (Hogg (1974), p. 915):

> *It should be observed, however, that if the outliers occur with the more interior values of the independent variables, their influence on least squares*

is minimal. On the other hand, outliers occurring with extreme values of the independent variables can be very disruptive.

Thus the priors favor $D_{\mathscr{S}}$ for the more disruptive subsets $\mathscr{S}$.

One final comment remains regarding what Pearson and Chandra Sekar (1941) and Murphy (1951) call the *masking effect.* Such an effect may occur when an outlier detector with $s = 1$ fails to reveal any outliers because of a mutual masking effect caused by multiple outliers in the data. Because the outlier problem has been shown equivalent to the selection of special dummy variables, this masking effect may be equated with the problems incurred in the stepwise inclusion of variables. Further discussion of the problems that occur in the sequential elimination of outliers is given in McMillan (1971) and McMillan and David (1971).

6. Slippage Tests. Closely related to outlier tests are the location and variance slippage tests. These tests are simply outlier detection tests where the possible outlying subsets are the strata of some fixed partition of the data. Such tests were first considered by Mosteller (1948), Paulson (1952), and Kudo (1956a) for location slippage, and by Cochran (1941) for variance slippage. Their formulations involve testing the equality of the strata populations with the alternative that one of the strata populations has *slipped* away from the others which remain identical.

These tests can be formulated as variable selection tests that use special independent dummy variables and allow only for the selection of certain variable combinations. For both location and variance slippage models, we assume that $\{1, \cdots, n_1\}$ indexes stratum 1 with each datum having mean $\boldsymbol{\mu}_1$ and covariance $\boldsymbol{\Sigma}_1 > \mathbf{0}$, $\{n_1 + 1, \cdots, n_1 + n_2\}$ indexes stratum 2 with mean $\boldsymbol{\mu}_2$ and covariance $\boldsymbol{\Sigma}_2 > \mathbf{0}$, etc. Let there be m strata of sizes $n_1, \cdots, n_m$ so that $\sum n_i = n$.

For location slippage, $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_m$ and the hypotheses of interest are

$$H_0: \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_m$$

$$H_i: \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_{i-1} = \boldsymbol{\mu}_{i+1} = \cdots = \boldsymbol{\mu}_m \neq \boldsymbol{\mu}_i \tag{6.1}$$

for $i = 1, \cdots, m$. Assume the model (2.1) with $\mathbf{X} = (1, \cdots, 1)'$ an $n \times 1$ location vector and $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m)'$, an $m \times k$ parameter matrix. Let $\mathbf{Z}$ consist of m column vectors such that the ith column is an indicator vector for the ith stratum. Then the variable selection tests in (3.2) and (3.3) with $s = 1$ result in the locational slippage tests proposed by Paulson (1952) and Karlin and Truax (1960).

COROLLARY 6.1. *Under model* (2.1) *and the multivariate locational slippage framework given above, the rule* (3.3) *where*

$$\Lambda_i = n_i(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})'\hat{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})/(n - n_i),$$

$$\bar{\mathbf{y}}_i = i\text{th } \textit{stratum mean}, \quad \bar{\mathbf{y}} = \textit{grand mean}, \tag{6.2}$$

and

$$(n-1)\hat{\boldsymbol{\Sigma}} = \textstyle\sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$$

is an admissible Bayes rule for the slippage problem in (6.1).

The balanced case of this test where $n_i = n/m = r$ was shown to be uniformly invariant admissible by Karlin and Truax (1960) and the same was shown for the balanced univariate case by Kudo (1956). Pfanzagl (1959) considered a locally optimum slippage test for the unbalanced univariate case. His treatment, however, was based on $\Lambda_i = n_i^2(\bar{y}_i - \bar{y})^2/\hat{\sigma}^2$ so it does not agree exactly with (6.2).

Slippage tests against alternatives which allow for S-strata slippage of equal (unequal) amount are possible by applying Theorem 3.1 with $s = 1$ ($s > 1$). Such tests have been proposed by Doornbos (1966) for the general case and by Ramachandran and Khatri (1957) when $s = 2$ and slippage occurs both upwards and downwards.

Theorem 4.1 allows the variance slippage problem to be considered in the balanced case where $n_i = n/m = r$. In this problem the hypotheses of interest are

$$(6.3)\qquad \begin{aligned} &H_0: \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_m \\ &H_i: \mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_{i-1} = (1+q)^{-1}\mathbf{\Sigma}_i = \mathbf{\Sigma}_{i+1} = \cdots = \mathbf{\Sigma}_m \end{aligned}$$

for $i = 1, \cdots, m$ where $q > 0$. We use model (2.1) with $\mathbf{X} = \mathbf{0}$, $\mathbf{\Delta} = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_n, \boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m)'$, and $\mathbf{Z} = (\mathbf{I}_n, \xi_1, \cdots, \xi_m)$, where ξ_i is an $n \times 1$ indicator variable for stratum i. The subsets from which we may choose include $(\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_r, \boldsymbol{\delta}_2, \cdots, \boldsymbol{\delta}_m)$, $(\boldsymbol{\eta}_{r+1}, \cdots, \boldsymbol{\eta}_{2r}, \boldsymbol{\delta}_1, \boldsymbol{\delta}_3, \cdots, \boldsymbol{\delta}_m)$, $\cdots$, $(\boldsymbol{\eta}_{(m-1)r}, \cdots, \boldsymbol{\eta}_{mr}, \boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_{m-1})$.

COROLLARY 6.2. *Under the variance slippage model described above, the test* (4.2) *with* $\Lambda_i = |\hat{\mathbf{\Sigma}}|/|\hat{\mathbf{\Sigma}}_i|$, *where*

$$\hat{\mathbf{\Sigma}} = \textit{within sum of squares over all strata}$$

and

$$\hat{\mathbf{\Sigma}}_i = \hat{\mathbf{\Sigma}} - \textit{within sum of squares of the } i\textit{th stratum}$$

is an invariant admissible Bayes rule for hypotheses (6.3).

PROOF. In order that the covariance matrix of each datum in a slipped stratum is the same, $\mathbf{Q}$ must be taken as

$$\mathbf{Q} = \begin{pmatrix} q\mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{pmatrix}$$

where $q > \mathbf{0}$ and $\mathbf{Q}_2 > \mathbf{0}$ is $(m-1) \times (m-1)$. The prior distribution on $\mathbf{Q}$ given in (4.8) may be used with $\mathbf{N} = (\mathbf{Z}_{\mathscr{S}}'\mathbf{M}\mathbf{Z}_{\mathscr{S}})^{-1/2}\mathbf{Q}^{-1/2}$. □

In the univariate case this test is the same as Cochran's (1941) test and is based on

$$(6.4)\qquad \max_i \hat{\sigma}_i^2 / \textstyle\sum_{j=1}^m \hat{\sigma}_j^2$$

where $\hat{\sigma}_i^2$ = error mean square for the ith stratum. The uniform invariant admissibility of (6.4) has been shown by Truax (1953).

The variance slippage test in Corollary 6.2 is invariant admissible for both location slippage and variance inflation. The justification of this is given in the previous section. Theorem 4.1 can also be used to construct variance slippage tests which allow for multiple strata slippage.

7. Concluding Remarks.

The choice of variables, outlier detection, and slippage problems that have been considered, assume that all k dimensions are involved in the selection process. If not, the information available in the uninvolved components should be added into the normal regression through the use of conditional densities. Specifically, let 1 and 2 index those portions of the data that are relevant and irrelevant in variable selection, so that

$$(\mathbf{Y}_1, \mathbf{Y}_2) = \mathbf{X}(\mathbf{\Theta}_1, \mathbf{\Theta}_2) + \mathbf{Z}(\mathbf{\Delta}, \mathbf{0}) + (\mathbf{E}_1, \mathbf{E}_2).$$

Then consider the likelihood in terms of the conditional distribution of $\mathbf{Y}_1$ given $\mathbf{Y}_2$ multiplied by the marginal distribution of $\mathbf{Y}_2$. Since the distribution of $\mathbf{Y}_2$ does not depend on $\mathbf{Z\Delta}$ then only the conditional density is relevant in the Bayesian development. This is equivalent to starting with the model

$$(7.1)\qquad \mathbf{Y}_1 = \mathbf{X}(\mathbf{\Theta}_1 - \mathbf{\Theta}_2\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}) + \mathbf{Y}_2\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} + \mathbf{Z\Delta} + \mathbf{E}_{1|2}$$

where

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$$

and the common row covariance of $\mathbf{E}_{1|2}$ is $\mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$. Model (7.1) is model (2.1) with $\mathbf{Y}_2$ incorporated into $\mathbf{X}$.

Finally, it seems that greater advantage could be taken of the equivalence of the three problems above. For example, Kapur (1957) has shown the unbiasedness of one-stratum slippage tests when more than one stratum has slipped. These ideas may be extendable to the more general problem of forward selection and backward elimination of variables.

Acknowledgments. This paper is based on part of the author's Ph.D. dissertation written at The University of Michigan under the supervision of Professor Edward D. Rothman. The author would like to thank Professors E. D. Rothman and Michael Woodroofe for their encouragement and help, and the referees for their format suggestions and patience.

APPENDIX

Formulae for updating a regression when including further independent variables

Transform $\mathbf{Y}$ by $\mathbf{V}$ so that

(A.1) $$\min_{\Theta, \Delta_{\mathscr{S}}} \operatorname{tr}(\mathbf{Y} - \mathbf{X}\mathbf{\Theta} - \mathbf{Z}_{\mathscr{S}}\mathbf{\Delta}_{\mathscr{S}})'(\mathbf{Y} - \mathbf{X}\mathbf{\Theta} - \mathbf{Z}_{\mathscr{S}}\mathbf{\Delta}_{\mathscr{S}})$$
$$= \min_{\Theta, \Delta_{\mathscr{S}}} \operatorname{tr} \textstyle\sum_{i=1}^{3} (\mathbf{U}_i - E\mathbf{U}_i)'(\mathbf{U}_i - E\mathbf{U}_i)$$
$$= \mathbf{U}_3'\mathbf{U}_3$$

by (3.7). This minimum occurs at $\mathbf{U}_1 = \boldsymbol{\mu}_{\mathscr{S}}$ and $\mathbf{U}_2 = \boldsymbol{\nu}_{\mathscr{S}}$. Then

(A.2) $$\mathbf{U}_3'\mathbf{U}_3 = \hat{\mathbf{S}}_{\mathscr{S}}, \qquad \hat{\mathbf{\Delta}}_{\mathscr{S}} = (\mathbf{Z}_{\mathscr{S}}'\mathbf{M}\mathbf{Z}_{\mathscr{S}})^{-1}\mathbf{Z}_{\mathscr{S}}'\mathbf{M}\mathbf{Y},$$

and

$$\hat{\mathbf{\Theta}}_{\mathscr{S}} = \hat{\mathbf{\Theta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}_{\mathscr{S}}\hat{\mathbf{\Delta}}_{\mathscr{S}}.$$

Also since $\sum \mathbf{U}_i'\mathbf{U}_i = \mathbf{Y}'\mathbf{Y}$, then

(A.3) $$\hat{\mathbf{S}}_{\mathscr{S}} = \mathbf{U}_3'\mathbf{U}_3 = \mathbf{Y}'\mathbf{Y} - \mathbf{U}_1'\mathbf{U}_1 - \mathbf{U}_2'\mathbf{U}_2 = \hat{\mathbf{S}} - \hat{\mathbf{E}}'\mathbf{Z}_{\mathscr{S}}(\mathbf{Z}_{\mathscr{S}}'\mathbf{M}\mathbf{Z}_{\mathscr{S}})^{-1}\mathbf{Z}_{\mathscr{S}}'\hat{\mathbf{E}}.$$

REFERENCES

ALLEN, D. M. (1971). The prediction sum of squares as a criterion for selecting prediction variables. Tech. Report No. 23, Dept. of Statistics, Univ. of Kentucky.

ALLEN, D. M. (1974). The relationships between variable selection and data augmentation and a method of prediction. *Technometrics* **16** 125–127.

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

BEALE, KENDALL and MANN (1967). The discarding of variables in multivariate analysis. *Biometrika* **54** 357–366.

BECKMAN and TRUSSEL (1974). The distribution of an arbitrary Studentized residual and the effects of updating in multiple regression. *J. Amer. Statist. Assoc.* **69** 199–201.

BOX, G. E. P. and TIAO, G. C. (1968). A Bayesian approach to some outlier problems. *Biometrika* **55** 119–129.

COCHRAN, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Ann. Eugenics* **11** 47–52.

DEMPSTER, A. P. (1969). *Elements of Continuous Multivariate Analysis*. Addison-Wesley, Reading.

DIXON, W. J. (1953). Processing data for outliers. *Biometrics* **9** 74–89.

DOORNBOS, R. (1966). *Slippage Tests*. Mathematical Centre Tracts, No. 15, Mathematisch Centrum, Amsterdam.

FERGUSON, T. S. (1960). On the rejection of outliers. Proc. Fourth Berkeley Symp. Math. Statist. Prob. **1** 253–287.

GENTLEMAN, J. F. and WILK, M. B. (1975). Detecting outliers, II. Supplementing the direct analyses of residuals. *Biometrics* **31** 387–410.

GRUBBS, F. E. (1950). Sample criteria for testing outlying observations. *Ann. Math. Statist.* **21** 27–58.

HOCKING, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32** 1-49.
HOGG, R. V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *J. Amer. Statist. Assoc.* **69** 909-928.
JEFFREYS, H. (1939). *Theory of Probability*. Oxford.
KAPUR, M. N. (1957). A property of the optimum solution suggested by Paulson for the k-sample slippage problem for the normal distribution. *Ind. Soc. Agri. Statist.* **9** 179-190.
KARLIN, S. and TRUAX, D. R. (1960). Slippage problems. *Ann. Math. Statist.* **31** 296-324.
KIEFER, J. and SCHWARTZ, R. (1965). Admissible Bayes character of T^2-, R^2-, and other fully invariant tests for classical multivariate normal problems. *Ann. Math. Statist.* **36** 747-770.
KUDO, A. (1956a). On the testing of outlying observations. *Sankhyā* **17** 67-76.
KUDO, A. (1956b). On the invariant multiple decision procedures. *Bull. Math. Statist.* **6** 57-68.
LAMOTTE, L. R. and HOCKING, R. R. (1970). Computational efficiency in the selection of regression variables. *Technometrics* **12** 83-93.
LINDLEY, D. V. (1968). The choice of variables in multiple regression. *J. Roy. Statist. Soc. Ser. B.* **30** 31-53.
MCMILLAN, R. G. (1971). Tests for one or two outliers in normal samples with unknown variance. *Technometrics* **49** 87-100.
MCMILLAN, R. G. and DAVID, H. A. (1971). Tests for one of two outliers in normal samples with known variance. *Technometrics* **13** 75-100.
MOSTELLER, F. (1948). A k-sample slippage test for an extreme population. *Ann. Math. Statist.* **19** 58-65.
MURPHY, R. B. (1951). On tests for outlying observations. Ph.D. thesis, Princeton Univ.
NEWTON, R. G. and SPURREL, O. J. (1967). A development of multiple regression for the analysis of routine data. *Appl. Statist.* **16** 51-65.
PAULSON, E. (1952). An optimum solution to the k-sample slippage problem for the normal distribution. *Ann. Math. Statist.* **23** 610-616.
PEARSON, E. S. and CHANDRA SEKAR, C. (1941). The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrika* **28** 308-320.
PFANZAGL, J. (1959). Ein kombiniertes test & klassifikations-problem. *Metrika* **2** 11-45.
QUESENBERRY, C. P. and DAVID, A. A. (1961). Some tests for outliers. *Biometrika* **48** 379-390.
RAMACHANDRAN, K. V. and KHATRI, C. G. (1957). On a decision procedure based on the Tukey statistic. *Ann. Math. Statist.* **28** 802-806.
SCHMIDT, P. (1973). Methods of choosing among alternative linear regression models. Univ. of N. Carolina.
SRIKANTAN, K. S. (1961). Testing for the single outlier in a regression model. *Sankhyā Ser. A* **23** 251-260.
STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B.* **36** 111-147.
THOMPSON, W. R. (1935). On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. *Ann. Math. Statist.* **6** 214-219.
TRUAX, D. R. (1953). An optimum slippage test for the variances of k normal distributions. *Ann. Math. Statist.* **24** 669-674.
WILKS, S. S. (1963). Multivariate statistical outliers. *Sankhyā Ser. A* **25** 407-426.

DEPARTMENT OF MATHEMATICS
THE UNIVERSITY OF TEXAS
AUSTIN, TX 78712