The Annals of Statistics
1990, Vol. 18, No. 2, 471–538

THE 1985 WALD MEMORIAL LECTURES

AN ANCILLARITY PARADOX WHICH APPEARS IN MULTIPLE LINEAR REGRESSION[1]

BY LAWRENCE D. BROWN

Cornell University

Consider a multiple linear regression in which Y_i, $i = 1, \dots, n$, are independent normal variables with variance σ^2 and $E(Y_i) = \alpha + V_i'\beta$, where $V_i \in \mathbb{R}^r$ and $\beta \in \mathbb{R}^r$. Let $\hat{\alpha}$ denote the usual least squares estimator of α. Suppose that V_i are themselves observations of independent multivariate normal random variables with mean 0 and known, nonsingular covariance matrix θ. Then $\hat{\alpha}$ is inadmissible under squared error loss if $r \geq 2$.

Several estimators dominating $\hat{\alpha}$ when $r \geq 3$ are presented. Analogous results are presented for the case where σ^2 or θ are unknown and some other generalizations are also considered. It is noted that some of these results for $r \geq 3$ appear in earlier papers of Baranchik and of Takada.

$\{V_i\}$ are ancillary statistics in the above setting. Hence admissibility of $\hat{\alpha}$ depends on the distribution of the ancillary statistics, since if $\{V_i\}$ is fixed instead of random, then $\hat{\alpha}$ is admissible. This fact contradicts a widely held notion about ancillary statistics; some interpretations and consequences of this paradox are briefly discussed.

1. Introduction. This paper introduces a general variety of admissibility paradox. It then continues with a more detailed study of this paradox as it operates in multiple linear regression. The paper concludes with some remarks about ancillary statistics. It is noted that the admissibility results of this paper contradict the widely held notion that statistical inference in the presence of ancillary statistics should be independent of the distribution of those ancillary statistics.

The general form of the paradox is presented in Section 2. The application to multiple linear regression is presented in Section 3 and some extensions of these results are presented in Section 4. Remarks about ancillary statistics are in Section 5.

Multiple linear regression. In multiple linear regression the dependent variables are assumed to be independent normal with mean a linear function of the vector V of predictor variables. The principal problem to be discussed is that of estimating the y-intercept value α, i.e., the population mean of the dependent variables when the predictor variables are all zero.

If the predictor variables take on any prespecified constant values (assuming only that α is estimable), then the least squares estimator $\hat{\alpha}$ is admissible

Received October 1987; revised December 1989.

[1] Research supported in part by NSF Grants DMS-85-06847 and DMS-88-09016.

AMS 1980 *subject classifications*. Primary 62C15, 62C20, 62F10, 62A99; secondary 62H12, 62J05.

Key words and phrases. Admissibility, ancillary statistics, multiple linear regression.

under squared error loss. However, if the predictor variables are independent normal vectors with mean **0** and known nonsingular covariance matrix, then $\hat{\alpha}$ is not admissible. Admissibility of $\hat{\alpha}$ thus depends on whether the ancillary statistic V has a degenerate distribution or a nondegenerate normal distribution.

Precursors: *Stein's paradox*. Probably the best known paradox in estimation theory was discovered by Charles Stein. It involves the simultaneous estimation of at least three normal means or other location parameters. [See Stein (1956) and James and Stein (1961), or see Berger (1985) for a contemporary survey.] The current paradox also involves at least three normal means or other location parameters. But in other respects the two paradoxes are structurally quite different. It is essential for Stein's paradox that one be interested in simultaneously estimating three (or more) parameters, whereas the current results involve estimation of just one. Also, ancillary statistics play no role in Stein's results, whereas here their presence is essential; and as has already been noted, the presence of the paradox depends on a certain feature of their distribution.

While the current paradox is structurally very distinct from Stein's, it is mathematically very closely related. From the mathematical perspective the current results are merely a slight variant of Stein's. Some of the current results, such as Theorem 2.1.2 and Lemma 3.3.3, are direct adaptations of the theory of Stein estimation to the current context. Other principal results of the paper, such as Theorem 2.2.1, Theorem 3.2.1, Lemma 3.3.4 and Lemma 3.3.5, etc., while not direct adaptations, are based on techniques of proof familiar in the theory of Stein estimation. The general theory in Section 2 and the organization of Section 3.3 have been planned so as to emphasize this close mathematical relation.

Two other remarks are relevant concerning the relation to Stein estimation:

On the structural side: Some of the improved estimators constructed in this paper are unbiased, whereas improved estimators of multivariate normal means must be biased. [Construction of unbiased dominating estimators for our problem requires that there be at least four (instead of three) unknown means. See Section 3.3 for some unbiased dominating estimators in the multiple regression problem.]

On the mathematical side: Stein (1960) considers the problem of prediction in multiple linear regression. The principal results there can be viewed as an adaptation of his earlier results in the multiple normal means problem. Most of the results in our Section 3.3 can then be viewed as a further adaptation of his results in prediction. Remark 2.1.3 discusses this mathematical connection. There does not appear to be a similar direct mathematical connection with respect to our results in Sections 2.2 and 3.2.

Priority: *Baranchik's prediction results*. Baranchik (1964, 1973) further develops the prediction formulation presented in Stein (1960). Several of the results in our paper are already present in Baranchik's papers. Our key

Lemma 3.3.1 appears as part of the proof of Theorem 2 of Baranchik (1973). His Theorem 2 involves only simultaneous estimation of both α and β, rather than estimation of α alone. However a referee has pointed out that Section 3.4 of the technical report [Baranchik (1964)] upon which the later paper was based, does present an inadmissibility result for α alone which is much like our Lemmas 3.3.1 and 3.3.5 combined. (Apparently Baranchik failed to realize the significance of this result and so omitted it from his later paper.)

Other precursors. There are some earlier, rather pathological, results in which admissibility of an estimator depends on the distribution of the ancillary statistic.

There is, first of all, a rather trivial observation concerning a one parameter location problem with ancillary statistic. (It holds in other problems as well.) The customary procedure can have everywhere finite risk conditional on the value of the ancillary statistic and yet have identically infinite risk unconditionally. Whether this occurs obviously depends in part on the distribution of the ancillary statistic. When this occurs the customary estimator is inadmissible.

Even when the unconditional risk is finite, the customary estimator can be inadmissible in the unusual situation that certain mild moment conditions are violated. This moment paradox was first observed by Brown (1966, page 1113) and Perng (1970). See also Fox (1981).

2. General theory.

2.1. *A simple paradigm.* Let $X \sim N(\mu, \Sigma)$, Σ known. Let $w \in \mathbb{R}^p$, with $\sum_{i=1}^p w_i^2 > 0$ and define

$$\theta = \sum_{i=1}^{p} w_i \mu_i = w'\mu. \tag{2.1.1}$$

Consider the problem of estimating θ under ordinary squared error loss

$$L(\mu, d) = (d - \theta)^2, \qquad d \in \mathbb{R}. \tag{2.1.2}$$

The customary estimator for this problem is defined by $\delta_0(x) = w'x$. As a preliminary result we note:

PROPOSITION 2.1.1. *In the problem formulated above, δ_0 is minimax and admissible.*

PROOF. This fact implicitly appears in Stein (1959) and explicitly appears in Cohen (1965), and was perhaps known as early as Blyth (1951) or Hodges and Lehmann (1951). □

Now assume that the values of $(w_1, \ldots, w_p)$ appearing in (2.1.1) are observed coordinate values of a random variable $W \in \mathbb{R}^p$. The simplest case occurs when the distribution of W is known and W is independent of X. This is the case treated in Theorem 2.1.2. More complex situations are discussed in the next subsection.

The customary estimator of θ in this problem remains $\delta_0(x, w) = w'x$. Let

$$\Omega = E(WW'). \tag{2.1.3}$$

(Assume Ω exists.)

THEOREM 2.1.2. *Let X, W be independent. Suppose the $p \times p$ matrix Ω is nonsingular and $p \geq 3$. Then δ_0 is not admissible for loss* (2.1.2). *A better estimate is given by $\delta^*(x, w) = w'd^*(x)$, where*

$$d^*(x) = x - \frac{\rho}{x'\Sigma^{-1}\Omega^{-1}\Sigma^{-1}x}\Omega^{-1}\Sigma^{-1}x \tag{2.1.4}$$

with $0 < \rho < 2(p - 2)$.

PROOF. Observe that for any estimator of the form $\delta(x, w) = w'd(x)$, $d \in \mathbb{R}^d$,

$$\begin{aligned} R(\mu, \delta) &= E_\mu(W'd(X) - W'\mu)^2 \\ &= E_\mu\{(d(X) - \mu)'WW'(d(X) - \mu)\} \\ &= E_\mu\{(d(X) - \mu)'\Omega(d(X) - \mu)\}. \end{aligned}$$

since W and X are independent. This is the same as the risk of $d(\cdot) \in \mathbb{R}^p$ under the loss $L(\mu, d) = (d - \mu)'\Omega(d - \mu)$. It is known that for this loss the estimator d^* is minimax and dominates $d(x) = x$. See Stein (1960) for the case $\Omega = cI$ and Berger (1976) or Hudson (1974) for the more general case. □

If Ω is singular but of rank greater than or equal to 3 a similar inadmissibility result is valid: Just use a generalized inverse of Ω in (2.1.4) and substitute the rank of Ω in place of p in that formula. If Ω has rank 1 or 2, then δ_0 is admissible by an extension of the reasoning in Proposition 2.1.1. [When the rank is 2 one also needs the two dimensional admissibility result in James and Stein (1961) or Brown (1971).]

REMARK 2.1.3. A situation somewhat analogous to the above has already been observed in the literature, first in Stein (1960). Consider a normal linear model (e.g., a multiple linear regression) of the form

$$Y \sim N_q(D\mu, \sigma^2 I). \tag{2.1.5}$$

Here $q \geq p$, D is a (known) design matrix of full rank, σ^2 is known and $\mu \in \mathbb{R}^p$ is the (unknown) parameter vector. Let $X = \hat{\mu} = (D'D)^{-1}D'Y$. Then $X \sim N_p(\mu, \Sigma)$ with $\Sigma = \sigma^2(D'D)^{-1}$.

Now suppose $W \in \mathbb{R}^p$ and $Z \in \mathbb{R}^1$ are future random variables independent of the vector Y and, given $W = w$, $Z \sim N(w'\mu, \sigma^2)$. Suppose it is desired to *predict* the value of Z after having observed X and W. Consider a prediction π

of the form $\pi(x, w) = w'd(x)$ and note that

$$\begin{aligned} E\big((\pi(X,W) - Z)^2\big) &= E\big((W'd(X) - W'\mu)^2\big) + \sigma^2 \\ &= L(\mu, W'd(X)) + \sigma^2 \end{aligned} \tag{2.1.6}$$

with L as defined in (2.1.2). When $p \geq 3$ inadmissibility of the usual prediction $\pi_0 = w'x$ under quadratic loss therefore follows from a result like Theorem 2.1.2. This result in prediction theory has more recently been noted and exploited in, e.g., Baranchik (1973), Copas (1983) and Oman (1984).

Lemma 3.31 can be viewed as an application of Theorem 2.1.2 which involves a multiple regression setup like that in (2.1.5); however, Theorem 2.1.2 is applied there in a quite different fashion than in the above example. In particular, throughout Section 3 there are no future observations W, Z under consideration and the problem is thus one of estimation rather than prediction.

Remark 2.1.4. The form of d^* in (2.1.4) is analytically convenient. However somewhat better estimates can be defined by substituting other forms for d^*. For example, one could use

$$d_+^*(x) = x - \min\left\{\max\operatorname{eig}(\Sigma Q), \frac{p}{x'\Sigma^{-1}\Omega^{-1}\Sigma^{-1}x}\right\}\Omega^{-1}\Sigma^{-1}x. \tag{2.1.7}$$

Sections 4.7.7, 4.7.10 and 5.4.3 of Berger (1985) contain a useful discussion of alternate estimators including many which, while not minimax, have other appealing properties.

2.2. *A more complex result.* In the simple setting of the previous subsection the covariance matrix Σ of X was assumed fixed and known. An inadmissibility result is also valid when Σ is random with W either random or fixed. This result is less satisfactory than the previous one in that we are unable here to provide a useful formula for an estimator which dominates δ_0. This defect is discussed below in more detail and also, indirectly, in the application of the next section.

For the following theorem let Q and Σ be observable $(p \times p)$ positive semidefinite matrix valued random variables. Assume Σ is positive definite. Let the joint distribution of (Q, Σ) be known and define

$$\Omega = E(\Sigma Q \Sigma). \tag{2.2.1}$$

(Assume it exists.) Suppose $X \sim N(\mu, \Sigma)$ and it is desired to estimate μ under the (random) loss

$$L(\mu, d) = (d - \mu)'Q(d - \mu). \tag{2.2.2}$$

The customary estimator here is $\delta_0(x) = x$.

To connect this formulation with that of the previous section assume Σ is fixed and let $Q = WW'$. Estimation of μ by $d \in \mathbb{R}^p$ under loss (2.2.2) is obviously equivalent to estimation of $\theta = W'\mu$ by $W'd$ under loss (2.1.2) since

$$(d - \mu)'Q(d - \mu) = (d - \mu)'WW'(d - \mu) = (W'd - W'\mu)^2 = (W'd - \theta)^2.$$

THEOREM 2.2.1. *Let $p \geq 3$. Suppose Ω defined by* (2.2.1) *is nonsingular. Then δ_0 is inadmissible under loss* (2.2.2).

PROOF. If $\Sigma = (\sigma_{ij})$ let $\|\Sigma\| = \max\{|\sigma_{ij}|: 1 \leq i, j \leq p\}$. Given $B < \infty$ let $\Omega_B = E(\Sigma Q \Sigma | \, \|\Sigma\| \leq B, \|Q\| \leq B)$. If $\Omega = \Omega_\infty$ is nonsingular, as hypothesized, there must be a $B < \infty$ such that Ω_B is nonsingular. Σ and Q are ancillary statistics; so it suffices to show that conditionally, given $\|\Sigma\| \leq B$ and $\|Q\| \leq B$, $\delta_0(x) = x$ is inadmissible. Accordingly we can now assume with no loss of generality that $\|\Sigma\| \leq B$ and $\|Q\| \leq B$ with probability 1. We will do so and omit explicit mention of B in the sequel.

Let $d > 0$ and

$$\delta(x, Q, \Sigma) = \left(I - \frac{\rho}{d + x'\Omega^{-1}x}\Sigma\Omega^{-1}\right)x, \qquad 0 < \rho < 2(p-2). \tag{2.2.3}$$

Applying Berger (1985, page 362) gives

$$\begin{aligned}\Delta &= R(\mu, \delta_0) - R(\mu, \delta) \\ &= 2\rho E\left\{E_\mu\left[\operatorname{tr}\left(\frac{\Omega^{-1}\Sigma Q \Sigma}{d + X'\Omega^{-1}X} - \left(2 + \frac{\rho}{2}\right)\frac{\Omega^{-1}XX'\Omega^{-1}\Sigma Q \Sigma}{(d + X'\Omega^{-1}X)^2}\right)\middle| Q, \Sigma\right]\right\}.\end{aligned}$$

Write $X = \mu + Z$, where $Z \sim N(0, \Sigma)$. Expanding the error term as in Brown (1966, pages 1122–1124) yields

$$\begin{aligned}&E_\mu\left[\operatorname{tr}\frac{\Omega^{-1}\Sigma Q\Sigma}{d + X'\Omega^{-1}X}\middle| Q, \Sigma\right] \\ &\quad= \operatorname{tr}\frac{\Omega^{-1}\Sigma Q \Sigma}{d + \mu'\Omega^{-1}\mu} E\left[\left(1 - \frac{2Z'\Omega^{-1}\mu + Z'\Omega^{-1}Z}{d + \mu'\Omega^{-1}\mu}\right.\right. \\ &\qquad\qquad \left.\left. + \frac{(2Z'\Omega^{-1}\mu + Z'\Omega^{-1}Z)^2}{(d + \mu'\Omega^{-1}\mu)(d + (Z + \mu)'\Omega^{-1}(Z + \mu))}\right)\middle| Q, \Sigma\right] \\ &\quad= \operatorname{tr}\left(\frac{\Omega^{-1}\Sigma Q \Sigma}{d + \mu'\Omega^{-1}\mu}\right) + O\left(\frac{1}{(d + \mu'\Omega^{-1}\mu)d}\right).\end{aligned} \tag{2.2.4}$$

The $O(\cdot)$ term above is uniform in Σ, Q since $\|\Sigma\| \leq B$ and $\|Q\| \leq B$ by assumption. [One basic inequality used in verifying (2.2.4) is

$$\frac{(z'\Omega^{-1}\mu + z'\Omega^{-1}z)^2}{d + (z + \mu)'\Omega^{-1}(z + \mu)} < \frac{d + \mu'\Omega^{-1}\mu}{d}.$$

This can be verified by examining the simple expression $s^2/(d + (s + t)^2)$ which is maximized for fixed, d, t by the choice $s = -(d + t^2)/t$.] Similar reasoning yields

$$E_\mu\left[\operatorname{tr}\frac{\Omega^{-1}XX'\Omega^{-1}\Sigma Q\Sigma}{(d + X'\Omega^{-1}X)^2}\middle| Q, \Sigma\right] = \operatorname{tr}\frac{\Omega^{-1}\mu\mu'\Sigma Q \Sigma}{(d + \mu'\Omega^{-1}\mu)^2} = O\left(\frac{1}{(d + \mu'\Omega^{-1}\mu)d}\right).$$

Taking expectations over Q, Σ now yields

$$\Delta = \frac{2\rho}{d + \mu'\Omega^{-1}\mu}\left(p - \left(2 + \frac{\rho}{2}\right)\operatorname{tr}\frac{\mu'\Omega^{-1}\mu}{d + \mu'\Omega^{-1}\mu}\right) + O\left(\frac{1}{(d + \mu'\Omega^{-1}\mu)d}\right)$$

$$\geq \frac{2\rho}{d + \mu'\Omega^{-1}\mu}\left(p - 2 - \frac{\rho}{2}\right) + O\left(\frac{1}{(d + \mu'\Omega^{-1}\mu)d}\right).$$

For any $0 < \rho < 2(p - 2)$ it is thus possible to choose sufficiently large d so that $\Delta > 0$ for all μ. □

Note that although δ does dominate δ_0, it is not shown above to be a practically useful competitor since d may need to be chosen to be quite large, in which case Δ is a quite small positive amount. The above proof does however suggest one should be able to find a dominating estimator satisfying

$$(2.2.5)\quad \delta(x, Q, \Sigma) = \left(I - \frac{\rho}{x'\Omega^{-1}x}\Sigma\Omega^{-1}\right)x + o(\|x\|^{-1}) \quad \text{as } \|x\| \to \infty,$$

$0 < \rho < 2(p - 2)$. It may often be that an estimator such as

$$(2.2.6)\qquad \delta(x, Q, \Sigma) = x - \min\left\{\min \operatorname{eig}(\Omega\Sigma^{-1}), \frac{p-2}{x'\Omega^{-1}x}\right\}\Sigma\Omega^{-1}x$$

dominates δ_0; however, results in Section 3.3 show that dominance of (2.2.6) can depend on the joint distribution of Q, Σ. (2.2.6) is, of course, the obvious, direct analog of d_+^* as defined in (2.1.7) with $\rho = p - 2$.

3. The multiple regression problem.

3.1. *Setting for fixed regression constants.* Consider the usual normal multiple linear regression. Denote the $(r + 1)$ unknown parameters by $\alpha \in \mathbb{R}$, $\beta = (\beta_1, \ldots, \beta_r)' \in \mathbb{R}^r$. Let $Y = (Y_1, \ldots, Y_n)'$ denote the observable random vector. Let $\sigma^2 > 0$ be a fixed, known constant. (See Section 4.1 for the case of unknown σ^2.) Let $V_i = (V_{i1}, \ldots, V_{ir})'$, $i = 1, \ldots, n$, denote the observed (i.e., known) regression constants. The coordinates $Y_1, \ldots, Y_n$ are assumed to be independent normal random variables with

$$(3.1.1)\qquad E(Y_i) = \alpha + V_i'\beta, \qquad \operatorname{Var}(Y_i) = \sigma^2, \qquad i = 1, \ldots, n.$$

One may alternately write $Y \sim N(\mathbf{1}\alpha + V\beta, \sigma^2 I)$ with $V = (v_{ij})$ and $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^n$. Assume V is of full rank. It is desired to estimate the y-intercept parameter α under ordinary quadratic loss:

$$(3.1.2)\qquad L((\alpha, \beta), d) = (d - \alpha)^2, \qquad d \in \mathbb{R}.$$

Some additional notation is needed in order to adequately describe the usual estimator. Let $\overline{Y} = n^{-1}\mathbf{1}'Y$, $\overline{V} = n^{-1}\mathbf{1}'V$ and $S = (V - \mathbf{1}\overline{V})'(V - \mathbf{1}\overline{V})$. [$\overline{Y}$ is a scalar, $\overline{V}$ is a $(1 \times r)$ *row* vector, and S is $(r \times r)$ and positive definite with probability 1.] The usual estimator of α (as well as the usual estimator of β) is the BLUE which is of course also the MVUE and MLE. It is given by

$$\begin{aligned} \delta_0 &= \hat{\alpha} = \bar{Y} - \bar{V}\hat{\beta}, \\ \hat{\beta} &= S^{-1}V'(Y - \bar{Y}\mathbf{1}). \end{aligned} \tag{3.1.3}$$

(Section 4.2 discusses estimation of linear contrasts other than α.)

Note that

$$\begin{aligned} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} &\sim N\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \Sigma(V) \right), \\ \Sigma(V) &= \sigma^2 \begin{pmatrix} n^{-1} + \bar{V}S^{-1}\bar{V}' & -\bar{V}S^{-1} \\ -S^{-1}\bar{V}' & S^{-1} \end{pmatrix}. \end{aligned} \tag{3.1.4}$$

Furthermore, $\bar{Y}$ is independent of $\hat{\beta}$.

Admissibility of $\hat{\alpha}$ is immediate from Proposition 2.1.1, as follows.

PROPOSITION 3.1.1. *In the preceding problem $\hat{\alpha}$ is an admissible estimator of α.*

PROOF. Let $X = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}$, $\mu = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ and $\Sigma = \Sigma(V)$ as in (3.1.4). Let $w = (1, 0, \dots, 0)'$. Then apply Proposition 2.1.1. □

3.2. *Random regression constants: Inadmissibility of $\hat{\alpha}$ for $r \geq 2$.*

Setting. In the preceding subsection the design variables v_{ij} were assumed to be fixed, known constants. This is realistic for applications where the v_{ij} are preset by the experimenter, perhaps in such a way that the experiment will satisfy some classical optimality criterion. [For example, one can choose the v_{ij} subject to $|v_{ij}| \leq B$ to minimize $\mathrm{tr}(\Sigma(V))$.] However, there are many other situations where the v_{ij} cannot be so closely controlled. Thus, in one broad class of situations the vectors $V_i = (v_{i1}, \dots, v_{ir})'$ are independent vector valued random variables.

The remainder of this paper (except for parts of Section 4) concerns the situation where the V_i are observations of independent random variables in $\mathbb{R}^r$ with a known distribution. Here, we take this known distribution to be normal with mean zero and covariance identity. (See Remark 3.2.1). Thus

$$V = \begin{pmatrix} V_1' \\ \vdots \\ V_n' \end{pmatrix}, \qquad V_i \sim N(0, I) \text{ (indep.)}, \qquad i = 1, \dots, n, \tag{3.2.1}$$

for given $V, Y \sim N(\mathbf{1}\alpha + V\beta, \sigma^2 I)$, σ^2 known. (See Section 4.1 for the case where σ^2 is unknown and estimable.)

REMARK 3.2.1. The distributional assumptions on V can be somewhat relaxed. Suppose that the V_i are independent with $V_i \sim N(0, \Theta)$, $i = 1, \dots, n$

(with Θ a known positive definite matrix), but otherwise (3.2.1) holds. Let $V_i^* = (\Theta')^{-1/2}V_i$ and $\beta^* = \Theta^{1/2}\beta$. Then (3.2.1) holds with V^*, β^* in place of V, β. Thus the results to follow apply after a simple transformation to this case. Various of the other distributional assumptions can also be relaxed without qualitatively changing the results to follow. Section 4 discusses further modifications of (3.2.1).

Inadmissibility. The usual estimator of α under quadratic loss (3.1.2) is still $\delta_0 = \hat{\alpha}$ as defined by (3.1.3). However, when $r \geq 2$ this estimator is now inadmissible, as the following theorem shows.

THEOREM 3.2.2. *Let V, Y be as in* (3.2.1). *Let $r \geq 2$. Then $\delta_0 = \hat{\alpha}$ is an inadmissible estimator of α under ordinary quadratic loss* (3.1.2).

PROOF. Let $\mu = \binom{\alpha}{\beta}$ and $\Sigma = \Sigma(V)$ as in (3.1.4) and the proof of Proposition 3.1.1. Here $\mu \in \mathbb{R}^p$ with $p = r + 1 \geq 3$. Note that now $V, \Sigma(V)$ are random variables with joint distribution specified via (3.2.1). Let $W = (1, 0, \ldots, 0)'$ and $Q = WW'$. As explained by (2.2.3), estimation of α under loss (3.1.2) is equivalent to estimation of μ under loss (2.1.2).

Suppose first that $n - r \geq 5$. Note that $\bar{V}$ and S are independent, $\bar{V} \sim N(0, I/n)$, $E(\bar{V}, \bar{V}') = I$ and $S \sim Wis(n - 1, I)$. Also, $E(S^{-2}) = kI$ with $k = (n - r - 2)^{-1}(n - r - 4)^{-1}(1 + (r - 1)(n - r - 1)^{-1})$ by Theorem 3.2(iii) of Haff (1979). Thus

$$\Omega = E(\Sigma(V)Q\Sigma(V))$$

$$= \sigma^4 E\begin{pmatrix} (n^{-1} + \bar{V}S^{-1}\bar{V}')^2 & -(n^{-1} + \bar{V}S^{-1}\bar{V}')\bar{V}S^{-1} \\ S^{-1}\bar{V}'(n^{-1} + \bar{V}S^{-1}\bar{V}') & S^{-1}\bar{V}\bar{V}'S^{-1} \end{pmatrix} \tag{3.2.2}$$

$$= \sigma^4\begin{pmatrix} n^{-2} + \dfrac{2r}{n^2(n - r - 3)} + \dfrac{r(r + 2)}{n^2(n - r - 3)(n - r - 5)} & \mathbf{0}_{(1\times r)} \\ \mathbf{0}_{(r\times 1)} & n^{-1}kI_{(r\times r)} \end{pmatrix}$$

since $(n(n - r - 1)/r)\bar{V}S^{-1}\bar{V}' \sim F_{r, n-p}$ and

$$(n^{-1} + \bar{V}S^{-1}\bar{V}')\bar{V}S^{-1} = -\left[(n^{-1} + (-\bar{V})S^{-1}(-\bar{V}'))(-\bar{V})S^{-1}\right].$$

Actually, the details of (3.2.2) are not important here; all that matters is that Ω is diagonal and hence nonsingular. It now follows from Theorem 2.2.1 that δ_0 is inadmissible.

If $n - r \leq 5$, then Ω, as defined above, does not exist. One must thus reason slightly differently. Let $S^{-1} = (s^{ij})$ and $\|S\| = \max\{|s^{ij}|: 1 \leq i, j \leq p\}$ and define $\|\bar{V}\| = \Sigma_{i=1}^p \bar{V}_i^2$ in the usual way. Let $B < \infty$. Consider the problem conditional on $\|S^{-1}\| \leq B$, $\|\bar{V}\| \leq B$. Consider $\Omega_B = E(\Sigma(V)Q\Sigma(V) | \|S^{-1}\| \leq B, \|\bar{V}\| \leq B)$. From (3.2.2) and symmetry considerations it is easy to see that Ω_B is a nonsingular diagonal matrix. Hence, from Theorem 2.2.1 it follows that δ_0

is inadmissible given $\|S^{-1}\| \le B$, $\|\bar{V}\| \le B$. Let δ' dominate δ_0 given this set; i.e.,

$$E\big((\delta' - \alpha)^2 \big| \|S^{-1}\| \le B, \|\bar{V}\| \le B\big) \le E\big((\delta_0 - \alpha)^2 \big| \|S^{-1}\| \le B, \|\bar{V}\| \le B\big)$$

with strict inequality for some parameter values. Define

$$\delta = \begin{cases} \delta', & \text{if } \|S^{-1}\| \le B, \|\bar{V}\| \le B, \\ \delta_0, & \text{otherwise.} \end{cases}$$

Then δ dominates δ_0; hence δ_0 is inadmissible. □

Theorems 3.2.2 and 2.2.1 do not provide a useful formula for an estimator which dominates $\delta_0 = \hat{\alpha}$. The discussion following Theorem 2.2.1 does, however, suggest the *conjecture* that $\hat{\alpha}$ is dominated by an estimator such as

$$(3.2.3) \qquad \delta = \hat{\alpha} - e'\left[\min\left\{\min \operatorname{eig}(\Omega\Sigma^{-1}), \frac{\rho}{x'\Omega^{-1}x}\right\}\Sigma\Omega^{-1}\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix}\right],$$

where $e' = (1, 0, \dots, 0) \in \mathbb{R}^{r+1}$ and $\Sigma = \Sigma(V)$ [see (3.1.4)], with $0 < \rho \le 2(p-2)$.

3.3. *Some minimax estimators for* $r \ge 3$. Throughout this section assume that the dependent and independent variables Y and V are both random with distribution given by (3.2.1).

A class of estimators. The preceding has shown that $\hat{\alpha}$, while minimax, is not admissible when $r \ge 2$. However the proof of that inadmissibility result is virtually nonconstructive—no useful alternative to $\hat{\alpha}$ is proven to dominate $\hat{\alpha}$. In this section we investigate a particular class of estimators to be defined by (3.3.1).

As of now the best of these estimators appears on the basis of partial evidence to be that given through formula (3.3.5). (See Section 3.4.) However, future research may alter this temporary conclusion in favor of an estimator qualitatively like that given through (3.3.6).

The reader interested only in the basic theory of the ancillarity paradox may skip directly to the general discussion in Section 5. (Perhaps Remark 3.4.3 and Section 4.6 would also be of interest in these general terms.)

The general class. Recall that $\hat{\alpha} = \bar{Y} - \bar{V}\hat{\beta}$. This expression motivates consideration of estimators of the form

$$(3.3.1) \qquad \tilde{\delta} = \bar{Y} - \bar{V}\tilde{\beta}(\hat{\beta}, S) = \hat{\alpha} + \bar{V}\big(\hat{\beta} - \tilde{\beta}(\hat{\beta}, S)\big),$$

where $\tilde{\beta}$ is a specified function of $\hat{\beta}$ and S only.

Use of (3.3.1) can also be motivated from invariance considerations under the transformations $(\alpha, \beta) \to (\alpha + a, M\beta)$ and $V \to M'V$ with $a \in \mathbb{R}$ and M an $(r \times r)$ orthogonal matrix.

There is another kind of motivational argument for (3.3.1) worth mentioning here. If it were known that $\beta = 0$, then $\bar{Y}$ would be the ordinary estimator of α and would be minimax and admissible. The estimator (3.3.1) with $\tilde{\beta}$ shrinking β toward 0 can thus the thought of as a smoothed pretest estimator. (An ordinary pretest estimator would be a discontinuous shrinker such as $\tilde{\beta} = 0$ if $\hat{\beta}'S\hat{\beta} < C$ and $\tilde{\beta} = \hat{\beta}$ otherwise.) The estimators involving $\tilde{\beta}_2$ and $\tilde{\beta}_3$ to be introduced later in this section can be given roughly such an interpretation; they can be written in the form $a\overline{Y} + (1 - a)\hat{\alpha}$ with $a = a(\hat{\beta}, S)$, where $a(\cdot)$ is decreasing as the data makes the null hypothesis $\beta = 0$ in some sense less likely. See Section 3.4 for further discussion of this. See Section 4.6 for a further motivation of (3.3.1) based on an empirical Bayes argument.

A basic lemma. The risk of the estimator in (3.3.1) compared to that of $\hat{\alpha}$ has a very simple form, as follows:

LEMMA 3.3.1. *Let $\tilde{\delta}$ be as in* (3.3.1). *Then*

$$\Delta = R((\alpha, \beta')', \hat{\alpha}) - R((\alpha, \beta')', \tilde{\delta}) \tag{3.3.2}$$
$$= n^{-1}\{E(\|\hat{\beta} - \beta\|^2) - E(\|\tilde{\beta} - \beta\|^2)\}.$$

PROOF. Let $\theta = \alpha + \overline{V}\beta = E(\overline{Y}|V)$. Note that $\overline{Y}$ is conditionally independent of $\hat{\beta}, S$ given $\overline{V}$ and that $\overline{V}$ is independent of $\hat{\beta}, S$. Hence

$$\begin{aligned} R((\alpha, \beta')', \tilde{\delta}) &= E(\|\overline{Y} - \overline{V}\tilde{\beta} - \theta + \overline{V}\beta\|^2) \\ &= E\Big((\overline{Y} - \theta)^2 - 2(\overline{Y} - \theta)\overline{V}(\tilde{\beta} - \beta) + (\tilde{\beta} - \beta)'\overline{V}'\overline{V}(\tilde{\beta} - \beta)\Big) \\ &= \sigma^2/n + E\big((\tilde{\beta} - \beta)'(I/n)(\tilde{\beta} - \beta)\big) \end{aligned}$$

since $E(\overline{Y} - \theta|\overline{V}, \tilde{\beta}) = E(\overline{Y} - \theta|\overline{V}, S) = 0$ and $E(\overline{V}'\overline{V}|\tilde{\beta}) = E(\overline{V}'\overline{V}|\hat{\beta}, S) = I/n$. The analogous expression is obviously valid for $\hat{\beta}$. Taking the difference yields (3.3.2). □

To summarize the significance of this result:

COROLLARY 3.3.2. *Admissibility of $\hat{\alpha}$ within the class of estimators of the form* (3.3.1) *is equivalent to admissibility of $\hat{\beta}$ as an estimator of β under ordinary quadratic loss* ($L = \|\hat{\beta} - \beta\|^2$) *within the class of estimators of the form $\tilde{\beta} = \tilde{\beta}(\hat{\beta}, S)$.*

Four varieties of minimax estimators of β. In view of Corollary 3.3.2, we now consider the problem of estimating β by $\tilde{\beta}(\hat{\beta}, S)$ under loss $L = n^{-1}\|\hat{\beta} - \beta\|^2$. Three special types of minimax estimators suggest themselves.

First, note that conditional on S the situation is exactly the classical one discussed in Section 2.1, with $\Sigma = \sigma^2 S^{-1}$ and $Q = n^{-1}I$. An estimator which, given S, conditionally dominates $\hat{\beta}$ also dominates $\hat{\beta}$ unconditionally. Consequently:

LEMMA 3.3.3. *The estimator*

$$\tilde{\beta}_1(\hat{\beta}, S) = \hat{\beta} - \min\left\{\min \operatorname{eig}(S^{-1}), \frac{\sigma^2 \rho}{\hat{\beta}' S^2 \hat{\beta}}\right\} S\hat{\beta} \tag{3.3.3}$$

dominates $\hat{\beta}$ *when* $0 < \rho \le 2(r-2)$ *and when* $n \ge r + 3$.

PROOF. The condition $n \ge r + 3$ is needed so that $E(\|\hat{\beta}\|^2) < \infty$. [See (3.4.1).] It also guarantees that S is nonsingular with probability 1. Note that the estimator (3.3.3) is just d_+^* of (2.1.7) so the lemma follows as in the proof of Theorem 2.1.2. (The customary choice for ρ would be $\rho = r - 2$.) □

The preceding estimator dominates $\hat{\beta}$ conditionally on S. Since S is random with known distribution (i.e., an ancillary statistic) it is plausible that even better estimators can be found which dominate $\hat{\beta}$ in expectation although they do not dominate conditionally given S. It is particularly tempting to consider estimators which do not depend on S. Here is such an estimator.

LEMMA 3.3.4. *Assume* $n \ge r + 3$. *The estimator*

$$\tilde{\beta}_2 = \left(1 - \min\left\{1, \frac{\sigma^2 \rho}{(n-r)\|\hat{\beta}\|^2}\right\}\right)\hat{\beta} \tag{3.3.4}$$

dominates $\hat{\beta}$ *when* $0 < \rho \le 2(r-2)$.

PROOF. Since $\tilde{\beta}_2$ is a function of $\hat{\beta}$ only, its risk depends only on the marginal distribution of $\hat{\beta}$. That marginal distribution is a multivariate t. This is of the form of Example 3 in Berger (1975), and the lemma follows immediately from Theorem 1 of that paper. □

There is another estimator which has been studied in connection with the prediction problem that is mathematically equivalent to the problem at hand. This estimator was first suggested by Stein (1960). This estimator is particularly natural in the variant of the current setting where $V_i \sim N(0, \Theta)$ and Θ is unknown; see Remark 3.2.1 and Section 4.4. The estimator is

$$\tilde{\beta}_3 = \left(1 - \min\left(1, \frac{\sigma^2 \rho}{\hat{\beta}' S \hat{\beta}}\right)\right)\hat{\beta}. \tag{3.3.5}$$

LEMMA 3.3.5. *Assume* $n \ge r + 3$. *The estimator* $\tilde{\beta}_3$ *dominates* $\hat{\beta}$ *when* $0 < \rho \le 2(r-2)$.

PROOF. This result is proved in Takada (1979) building on earlier results of Baranchik (1973). The proof there is explicitly for the unknown σ^2 case (as in Section 4.1) but is easily adapted to the known σ^2 case. □

In summary, the estimators $\tilde{\delta}_i$ constructed from $\tilde{\beta}_i$ via the recipe (3.3.1), $i = 1, 2, 3$, all dominate $\hat{\alpha}$ when $r \geq 3$, $n \geq r - 3$ and $0 < \rho \leq 2(r - 2)$.

REMARK 3.3.6. It can be shown that all admissible estimators of β in the setting of Corollary 3.3.2 must be generalized Bayes with respect to a prior over $\{\beta \in \mathbb{R}^r\}$. Hence it would seem desirable to use an estimator $\tilde{\beta}$ which is generalized Bayes or at least closely mimics the behavior of a generalized Bayes estimator. This is discussed in Brown (1987). In a sense made precise there it is shown that none of $\tilde{\beta}_1$, $\tilde{\beta}_2$ or $\tilde{\beta}_3$ has this desirable property. The desire to mimic the behavior of a generalized Bayes estimator suggests use of an estimator such as

$$\tilde{\beta}_5 = \left[I - \min\left\{ \min \operatorname{eig}(S), \frac{\sigma^2 \rho}{\|\hat{\beta}\|^2} \right\} S^{-1} \right] \hat{\beta} \tag{3.3.6}$$

which is just (2.2.6) rewritten for this special case. [Formula (4.6.3) presents another possibility.] We have not been able to analytically manage the risk of this estimator. However we have been able to verify that

$$\tilde{\beta}_4 = \left[I - \frac{\sigma^2 \rho S^{-1}}{\|\hat{\beta}\|^2} \right] \hat{\beta} \tag{3.3.7}$$

dominates $\hat{\beta}$ when $r \geq 3$ and n is sufficiently large. [See Brown (1987) for a precise statement.] However n needs to be moderately large before this domination occurs, and it does not appear that (3.3.7) provides a good solution to the problem of estimating β.

3.4. *Improvement of $\tilde{\delta}$ over δ_0.* The positive part James–Stein estimator δ_+ for the classical problem of Section 2.1 with $Q = I = \Sigma$ can produce quite dramatic improvement in risk. For that problem $R(\mu, \delta_0) = p$ and $\inf_{\mu \in \mathbb{R}^p} R(\mu, \delta_+) = R(0, \delta_+) < 2$. Hence the proportional improvement in risk can be larger than $1 - 2/p$, a quite dramatic amount when p is large. As has been frequently pointed out (and will be mentioned again below) this proportional improvement decreases to 0 quite rapidly as $\|\mu\|$ increases.

For moderately small values of n the estimators δ of Section 3.3. yield similar dramatic maximum improvements in risk. They do not do so for larger values of n. Note that via a simple direct calculation

$$\begin{aligned} R(\alpha, 0; \delta_0) &= \sigma^2 n^{-1}(1 + E(\operatorname{tr} S^{-1})) \\ &= \begin{cases} \sigma^2 n^{-1}\left(1 + \dfrac{r}{n - r - 2}\right), & \text{if } n \geq r + 3, \\ \infty, & \text{if } n \leq r + 2. \end{cases} \end{aligned} \tag{3.4.1}$$

For $\tilde{\delta}_i$ constructed via (3.3.1) from $\tilde{\beta}_i$, $i = 1, 2, 3$, here are some corresponding formulas for $n \geq r + 3$:

$$R(\alpha, 0; \delta_0) - R(\alpha, 0; \tilde{\delta}_1) > \left(\frac{\sigma^2}{n}\right) E_{\beta = 0}\left(\frac{2\rho(r - 2) - \rho^2}{\hat{\beta}' S^2 \hat{\beta}/\sigma^2}\right) \tag{3.4.2}$$

[see Berger (1985, page 364)];

$$(3.4.3) \qquad R(\alpha,0;\delta_0) - R(\alpha,0;\tilde{\delta}_2) > \frac{\sigma^2}{n(n-r)}\left(2\rho - \frac{\sigma^2}{(r-2)}\right),$$

so when $\rho = r - 2$,

$$(3.4.3') \qquad R(\alpha,0;\delta_0) - R(\alpha,0;\tilde{\delta}_2) > \frac{\sigma^2(r-2)}{n(n-r)}$$

[see Berger (1975)];

$$(3.4.4) \quad R(\alpha,0;\delta_0) - R(\alpha,0;\tilde{\delta}_3) > \frac{\sigma^2}{n(n-r-2)}(2\rho(r-2) - \rho^2)$$

[see Stein (1960)], so when $\rho = r - 2$

$$(3.4.4') \qquad R(\alpha,0;\delta_0) - R(\alpha,0;\tilde{\delta}_3) = \frac{\sigma^2(r-2)}{n(n-r-2)}.$$

For $\tilde{\delta}_3$ the proportional improvement in risk at $(\alpha, 0)'$ is

$$(3.4.5) \qquad \frac{R(\alpha,0;\delta_0) - R(\alpha,0;\tilde{\delta}_3)}{R(\alpha,0;\delta_0)} > \frac{r-2}{n-2}.$$

While this can approach 1 [for $r \to \infty$ and $(n - r)/r \to 0$], it is ordinarily much less. For example, for $n = 20$, $r = 3, 5, 10$ (resp.) it is 0.06, 0.17, 0.44 (resp.).

REMARK 3.4.1. The right sides of (3.4.2)–(3.4.5) are values for the estimators corresponding to $\tilde{\delta}_i$, $i = 1, 2, 3$, without the positive part adjustment which has been included in their definition. It is easy to show that the positive part adjustment decreases the risk at $(\alpha, 0)'$ and, hence, yields strict inequality in (3.4.2)–(3.4.4). It is also true that *the risk at* $(\alpha, 0)'$ *of the given positive part estimators is strictly decreasing in* ρ. Hence the right sides of (3.4.3′) and (3.4.4′) are lower bounds for improvement in risk when $\rho \geq r - 2$.

REMARK 3.4.2 (A summary conclusion). On the basis of this evidence it appears that the estimator of choice, *at this stage of research*, is $\tilde{\delta}_3$ as in (3.3.1) and (3.3.5), with $\tilde{\delta}_2$ a close competitor. Of course for a more definitive comparison one should really compare the functions $R(\alpha, \beta; \tilde{\delta}_i)$, $i = 1, 2, 3$, for various values of $\|\beta\|$ rather than just for $\beta = 0$ as has been done above. A numerical study could feasibly yield such a comparison. It should, however, be borne in mind as noted in Remark 3.3.6 that future research may discover an estimator based on an expression like (3.3.6) which is preferable to $\tilde{\delta}_I$, $i = 1, 2, 3$.

Note that $\tilde{\delta}_3$ has the additional advantage of being robust for misspecification of Θ, the covariance matrix of the V_i (see Section 4.4.) As a secondary observation, note that $\tilde{\delta}_2$ might appear preferable to $\tilde{\delta}_3$ because of its somewhat simpler form; however, this should generally not be a significant factor since the matrix S will ordinarily already be available from the computation of $\hat{\alpha}, \hat{\beta}$.

REMARK 3.4.3 (An argument in favor of $\tilde{\delta}$). While proportional gains in risk in the range of 0.06 to 0.44 as noted following (3.4.5) are not as dramatic as the figure $(1 - 2/p)$ available from the classical James–Stein situation, they seem large enough to be often worth seeking.

Of course it is the case here (as with the James–Stein problem) that the proportional gain in risk decreases rapidly to zero as $\|\beta\|$ increases. However, there is a factor which makes this decrease for larger $\|\beta\|$ often of less relevance than it is in the James–Stein situation. Explanation of that factor is the main goal of this remark.

The results of Section 3 have concerned regression problems in which interest is in the unadjusted mean α. *In such problems the regression parameters are often thought a priori to be near zero.* [They may for example correspond to measurements on concomitant variables suspected to have little or no influence on the measurement (Y) of primary interest; data which were collected, perhaps, only because it was convenient to do so and which once collected cannot easily be totally ignored.] Such situations are familiar in practice, and one frequent suggestion is to use a preliminary test estimator—using $\delta = \hat{\alpha}$ if the test of $\|\beta\| = 0$ rejects and $\delta = \bar{Y}$ if it accepts. As has already been explained, some of the estimators of Section 3.3 can be viewed as smoothed preliminary test estimators. In addition, their use can be justified on the basis of minimaxity in a way that (it seems) preliminary test estimators cannot be, and one may expect from other research on preliminary test estimators that the estimators of Section 3.3, or something much like them, should behave better than any preliminary test estimator. See, e.g., Sclove, Morris, and Radhakrishnan (1972) and Judge and Bock (1978, Chapter 12).

There are situations in which only some of the regression parameters $\beta_1, \dots, \beta_r$ are suspected a priori of being near zero. The considerations of Section 3.3 can be easily modified for such a situation by using $\hat{\beta}_i$ in (3.3.1) to estimate parameters not suspected of being near zero and using a vector estimator like $\tilde{\beta}$ for those suspected of being near zero. If at least three parameters are suspected of being near zero, then one can in this way improve on the usual estimator $\delta_0 = \hat{\alpha}$. (Some improvement is also possible, as in Section 3.1 when just two parameters are suspected of being near zero.)

4. Variants of the multiple regression problem.

4.1. *Unknown error variance.* In the most common applications the error variance σ^2 is actually unknown, but is estimable. So, suppose in (3.1.1) that σ^2 is unknown but there is available an independent variable U with U/σ^2 distributed as χ^2_m. Usually $m = n - r - 1$.

THEOREM 4.1.1. *Let* $r \geq 3$ *in the above setting. Replace* σ^2 *in the definitions* (3.3.8)–(3.3.10) *of* $\tilde{\beta}_i$, $i = 1, 2, 3$, *by* $U/(m + 2)$. *Then the resulting* *estimators of β remain minimax when $0 < \rho \le 2(p-2)$ and the corresponding estimators* (3.3.1) *of α are also minimax and dominate $\hat{\alpha}$.*

PROOF. The statement of $\tilde{\beta}_1$ follows immediately from Berger (1976) as did the proof of Lemma 3.3.3. The proofs for the modified $\tilde{\beta}_2$ and $\tilde{\beta}_3$ result from substituting $U/(m+2)$ for σ^2 in the original proofs, taking the additional expectation in the appropriate risk expressions for $\tilde{\beta}_2$ and $\tilde{\beta}_3$ [see Berger (1975)] and performing some simple, straightforward algebra. The details are thus omitted. □

The above result also holds for the estimators to be introduced in Sections 4.2 and 4.4. In much the same way it is possible to prove an extension of Theorems 2.2.1 and 3.2.2.

THEOREM 4.1.2. *$\hat{\alpha}$ is an inadmissible estimator in the above setting when $r = 2$ and $m \ge 5$ or when $r \ge 3$.*

PROOF. The result for $r \ge 3$ is just Theorem 4.1.1. For $r = 2$ write $\overline{\Sigma} = \Sigma(V)/\sigma^2$ with $\Sigma(V)$ as in (3.1.4) and $\overline{\Omega} = \Omega/\sigma^4$ with Ω defined by (3.2.2). Then let

$$\tilde{\delta} = \left[I - \frac{\rho}{d + U/m + x'\overline{\Omega}^{-1}x}\overline{\Sigma}\overline{\Omega}^{-1}\right]x. \tag{4.1.1}$$

Now proceed as in the proof of Theorem 2.2.1 to show that

$$\begin{aligned}\Delta &\ge E\left(\frac{2\tau\sigma^2}{d + U/m + \mu'\overline{\Omega}^{-1}\mu}\right)\left(p - 2 - \frac{\rho}{2}\right)\\ &\quad + O\left[E\left(\frac{\sigma^2}{(d + U/m + \mu'\overline{\Omega}^{-1}\mu)d}\right)\right],\end{aligned} \tag{4.1.2}$$

uniformly in μ, σ^2 as $d \to \infty$. In verifying this it is important that

$$\begin{aligned}&E\left(\frac{Z'\overline{\Omega}^{-1}Z}{(d + U/m + \mu'\overline{\Omega}^{-1}\mu)^2}\right)\\ &\quad = O\left[E\left(\frac{1}{(d + U/m + \mu'\overline{\Omega}^{-1}\mu)d}\right)\right]\end{aligned} \tag{4.1.3}$$

with $Z \sim N(0, \Sigma(V))$. The condition $m \ge 5$ guarantees that $E(\sigma^4/U^2)$ is bounded, which then yields (4.1.3). The theorem now follows from (4.1.2). [We suspect that the condition $m \ge 5$ is not required for validity of the theorem, although it does seem to be necessary for (4.1.3).] □

4.2. *Estimation of other contrasts.* The preceding has concentrated entirely on estimation of the single (linear) contrast α. It may instead be desired to estimate some other linear contrast. Denote such a contrast as $\kappa = a\alpha + b'\beta$

with $a \in \mathbb{R}$ and $b \in \mathbb{R}^r$ known and not both zero. The existence theory of Section 2.3 yields the following parallel to Theorem 3.2.

THEOREM 4.2. *Let V, Y be as in* (3.2.1). *Let $r \geq 2$. Then $\delta_0 = a\hat{\alpha} + b'\hat{\beta}$ is an inadmissible estimator of κ under ordinary quadratic loss $L = (\delta - \kappa)^2$.*

PROOF. The proof is parallel to that of Theorem 3.2.2. The matrix Q is now taken to be $Q = \binom{a}{b}(a, b')$. The resulting matrix $\Omega = E(\Sigma(V)Q\Sigma(V))$, when it exists, is still nonsingular, although it is of course no longer given by (3.2.2). When Ω as above does not exist one must first condition on a subset of $\{V\}$ just as in the proof of Theorem 3.2.2. The theorem thus follows from Theorem 2.2.1 as did Theorem 3.2.2. □

As before, the preceding does not yield a useful formula for an alternative to δ_0, although one can again conjecture that something rather like (2.2.6) and (3.2.3) will eventually be shown to yield a practically viable alternative to δ_0. When $r \geq 4$ and $n \geq r + 4$ it is possible to produce a useful alternative estimator dominating δ_0 which involves an expression like $\tilde{\beta}_i$, $i = 1, 2$ or 3, of Section 3.3. The derivation is somewhat lengthy and not entirely satisfactory. Details appear in Brown (1987, pages 53–56).

4.3. *Different mean for V.* To now we have assumed known that $E(V_i) = 0$. If it is known that $E(V_i) = \nu$ with $\nu \neq 0$, then the general pattern of inadmissibility revealed in Section 3.2 remains valid but the results of Section 3.3 seem to apply only through the methodology of Section 4.2.

Note that if $V_i \sim N(\nu, I)$ (independent) and [as in (3.2.1)] $Y \sim N(\mathbf{1}\alpha + V\beta, \sigma^2 I)$, σ^2 known, then

$$(4.3.1) \qquad Y \sim N\big(\mathbf{1}(\alpha + \nu'\beta) + (V - \mathbf{1}\nu')\beta, \sigma^2 I\big) = N\big(\alpha^* + V^*\beta, \sigma^2 I\big), \quad \text{say.}$$

Here, the rows of V^* are independent $N(0, I)$ variables, as in (3.2.1). It should be clear that estimation of α on the basis of V, Y as in (4.3.1) is equivalent to estimation of the contrast $\alpha^* - \nu'\beta$ in the setting of (3.2.1). The following thus summarizes the preceding results as they now apply.

THEOREM 4.3.1. *Let $E(V_i) = \nu$ (ν known) as in* (4.3.1) *above. Let $r \geq 2$. Then $\delta_0 = \hat{\alpha}$ is an inadmissible estimator of α under ordinary quadratic loss.*

If $r \geq 4$ and $n \geq r + 4$ an estimator improving on δ_0 can be found by proceeding as in the remark at the end of Section 4.2 to estimate the contrast $\alpha^* - \nu'\beta$ on the basis of (V^*, Y). Estimation of contrasts other than α can be discussed similarly.

REMARK 4.3.2. In most applications V, Y are as in (4.3.1) but ν is unknown. We *conjecture* that $\delta_0 = \hat{\alpha}$ is then an admissible estimator of α. In certain other applications ν is unknown but there also exist additional independent observations on V unaccompanied by a corresponding Y. These could be used to produce an independent estimator of ν. It can be shown under suitable conditions that δ_0 is then inadmissible when $r \geq 2$.

4.4. *Unknown covariance matrix of* V. Remark 3.2.1 has already discussed the situation where $V_i \sim N(0, \Theta)$ (independent) with Θ a known positive definite matrix. (Strictly speaking, when Θ is unknown, then S is no longer an ancillary statistic since its distribution depends on the unknown nuisance parameter Θ.) Observe that if Θ is unknown it can be estimated by $S/(n-1)$. Bearing this in mind along with the general nature of the problem it is natural to *conjecture* that $\hat{\alpha}$ is inadmissible if $r \geq 2$.

This conjecture is certainly valid when $r \geq 3$. The following result is now easy to establish and so its proof is omitted. We also note that the following result reformulated as a prediction problem is essentially what was stated and proved in Baranchik (1964, Lemma 3.1). [See also Takada (1979)].

THEOREM 4.4.1. *The risk of* $\tilde{\delta}_3$ *as an estimate of* α *is independent of* Θ (*so long as* Θ *is nonsingular*). *Consequently*, $\tilde{\delta}_3$ *dominates* $\hat{\alpha}$ *for all* (*nonsingular*) Θ *whenever* $0 < \rho \leq 2(r-2)$.

There is an interesting contrast between the result of Theorem 4.4.1 and the results and conjectures of Section 4.3. Note that nothing need be known about the covariance matrix Θ of V in order for $\hat{\alpha}$ to be inadmissible, but it appears that one does need information about the mean vector ν for V. Now, when $\nu = 0$ and $\Theta = I$ is known one might expect that knowledge of Θ should be used and would lead to a noticeably better estimator. Improvement of the estimators $\tilde{\delta}_1$, $\tilde{\delta}_2$ and $\tilde{\delta}_4$ over $\hat{\alpha}$ does depend on Θ. However, it does not appear that any of these estimators is to be preferred over $\tilde{\delta}_3$. (Indeed, the reverse is probably true, as noted in Remark 3.4.2.) However, as noted in Remark 3.3.6, it should be that an estimator somewhat like $\tilde{\delta}_5$ [derived from $\tilde{\beta}_5$ of (3.3.6)] should dominate $\tilde{\delta}_3$ when $\Theta = I$. It remains to be seen whether such dominance would be by a numerically significant amount.

4.5. *Other variants*. Other assumptions can undoubtedly be relaxed without altering the basic fact that the usual estimator of the single parameter α is inadmissible. Of course, alteration of these assumptions may affect the specific formulae [such as (3.3.1) and (3.3.3)–(3.3.6)] for improved estimators. Prime candidates for modification in future research are the assumptions that V be normal and that the residuals $Y - E(Y|V)$ be normal, along with the assumption of quadratic loss in the estimation of α (or some other contrast).

4.6. *An empirical Bayes interpretation*. Suppose the parameters (α, β) have a (formal) prior distribution under which α has a uniform prior and β has an (independent) normal prior with mean 0 and covariance $\tau^2 I$. Then the (formal) Bayes estimator of α is the (formal) posterior mean

$$\tilde{\delta} = \hat{\alpha} + \sigma^2 \overline{V} S^{-1} \left(\sigma^2 S^{-1} + \tau^2 I \right)^{-1} \hat{\beta}. \tag{4.6.1}$$

Note that

$$(4.6.2) \qquad E(\|\hat{\beta}\|^2 | S) = \|\beta\|^2 + \operatorname{tr} S^{-1} \quad \text{and} \quad E(\|\beta\|^2) = \tau^2 r.$$

A crude estimate of τ^2 is therefore $\|\hat{\beta}\|^2/r$. Substituting this in (4.6.1) yields the empirical Bayes estimate

$$(4.6.3) \qquad \tilde{\delta}_{\mathrm{EB}} = \hat{\alpha} + \sigma^2 \overline{V} S^{-1} \left(\frac{\sigma^2 r}{\|\hat{\beta}\|^2} S^{-1} + I \right)^{-1} \frac{r \hat{\beta}}{\|\hat{\beta}\|^2}.$$

The factor $(\sigma^2 r / \|\hat{\beta}\|^2) S^{-1}$ is nearly negligible when $\|\hat{\beta}\|^2$ is large. If this factor is omitted from (4.6.3) the resulting estimator is exactly of the form $\tilde{\delta}_4$ given by (3.3.1) and (3.3.6) with $\rho = r$. The estimator (4.6.3), or a modification of it using a less crude estimate of τ^2 (and possibly a constant ρ in place of the second r in the formula), is thus another good candidate as a minimax estimator to improve on $\delta_0 = \hat{\alpha}$ when $\theta = I$.

5. Ancillarity. The admissibility results of this paper are paradoxical in two ways. The first and more obvious one is that they show to be inadmissible the intuitively appealing estimator $\hat{\alpha}$—an estimator which also happens to be best invariant and minimax. The second and possibly more significant paradox is that they apparently contradict a widely held belief about the role of ancillary statistics. This section briefly discusses the ancillarity paradox.

The role of ancillary statistics. An ancillary statistic is one whose distribution is independent of the parameters of the statistical problem. It is widely held that statistical inference should be carried out conditional on the value of any ancillary statistic. That is to say, one should proceed as if the observed value of the ancillary statistic were a fixed, known constant. Another way of phrasing this is to say that the distribution of the ancillary statistic should be irrelevant to the statistical inference.

Let me quote briefly from Savage's (1976) paper, "On rereading R. A. Fisher." This quotation refers explicitly to the regression setting, described above, in which the value of the independent variable(s) is an ancillary statistic. [The reference in the quotation is actually only to the case of inference about the regression coefficient in ordinary linear regression (= one independent variable), but there is no reason to think that Savage or Fisher would have considered the multivariate linear regression problem studied here to be any different in this regard.]

"Fisher believes, and most of us with him, that if the statistic is ancillary, inference can be made from the conditional distribution of the data, given the parameters of interest and the ancillary statistic. That, for example, is how everyone ordinarily studies the regression coefficient. . . .

. . . the conclusion of statisticians of all persuasions has seemed to be that the conditional distribution of the regression coefficient given the value of the [independent variables] is appropriate for inference."

(Savage's discussion continues with a Bayesian interpretation for the belief described here.)

Cox's example. Cox (1958) has presented an important example relating to conditioning on an ancillary statistic. This example involves a hypothesis testing problem. The ancillary statistic is the sample size, which is either $n_1 = 1/\sigma_1^2$ or $n_2 = 1/\sigma_2^2$, $n_1 \ll n_2$. The formally optimal level $\alpha = 0.05$ text has conditional level nearly 0.1 if n_1 obtains and nearly 0 if n_2 obtains. Cox concludes:

"Now if the object of the analysis is to make statements by a rule with certain specified long-run properties, the unconditional test just given is in order, although it may be doubted whether the specification of desired properties is in this case very sensible. If, however, our object is to say 'what we can learn from the data that we have', the unconditional test is surely no good. Suppose that we know we have an observation from Σ_1. The unconditional test says that we can assign this a higher level of significance than we ordinarily do, because if we were to repeat the experiment, we might sample some quite different distribution. But this fact seems irrelevant to the interpretation of an observation which we know came from a distribution with variance σ_1^2. That is, our calculations of power, etc., should be made conditionally within the distribution known to have been sampled, i.e., if we are using tests of the conventional type, the conditional test should be chosen."

"To sum up, ... information as to whether it was Σ_1 [corresponding to n_1] or Σ_2 [corresponding to n_2] that we sampled tells us nothing about θ, and hence we make our inference conditionally on Σ_1 or Σ_2."

Thus, Cox's paradox indicates that the classical formulation is inappropriate for the testing problem he has considered. Brown (1978), following Kiefer (1976, 1977), argues similarly for a class of problems including that of Cox, and presents a formulation to supplement the classical one.

One can ask a similar question about the ancillarity paradox presented in the current article: does the paradox suggest that the classical decision theoretic formulation is inappropriate and needs to be altered? I think the answer is "No".

Cox's article makes a distinction between tests and confidence procedures on the one hand and point estimators on the other. Note that tests and confidence procedures involve a terminal decision statement plus a stochastic claim as to the accuracy of the terminal statement. For *validity* of the overall procedure that stochastic claim must be useful and correct. This is what breaks down in Cox's paradox. The stochastic claim that the unconditional level is $\alpha = 0.05$ is not useful. In fact, to say that the test performed has level $\alpha = 0.05$ is nearly certain to be interpreted in practice as a claim that the test performed has conditional level $\alpha = 0.05$ given the structure of the data actually observed; and such a claim is invalid.

In point estimation there is no such difficulty. An estimate may be nearer or further from the true value but (unless also accompanied by a confidence procedure) there is no statement concerning how close to true the estimate is. The only warranty the statistician can give is that he has done his best in the sense of providing an admissible (or nearly admissible) procedure which is also reasonable in the face of whatever generally acceptable a priori evaluations can be made about the parameter. Since no conditionally interpretable stochastic claim is being made, the estimation procedures discussed in this paper are conditionally *valid* and there thus seems no reason not to consider their admissibility in the classical (unconditional) formulation.

Conditional and unconditional admissibility. Ordinary notions of consistency demand use of procedures which are valid and admissible both conditionally and unconditionally. (Numerically minor deviations from this goal may be satisfactory and justifiable on the grounds of convenience. The preceding statement also requires the qualification that the problem be correctly modelled, otherwise it may be desirable to adopt robust but formally inadmissible procedures to reflect realistic possibilities that have been omitted from the formal model.)

Under normal circumstances Bayes procedures for ordinary (proper) priors attain this goal. [Even here there is a technical qualification which must be added to avoid pathologies. The unconditional (= marginal) Bayes risk must be finite. It is possible to formally construct statistical examples in which an ordinary Bayes procedure always has finite posterior risk given the data and even finite expected posterior risk given the value of an ancillary statistic and yet has infinite expected risk marginally and is inadmissible.]

The regression example in Section 3 shows that estimators (such as $\hat{\alpha}$) which are formally Bayes with respect to prior measures having infinite mass may easily be conditionally admissible and yet unconditionally inadmissible. Consider an estimator such as $\tilde{\delta}_1$. This estimator or one qualitatively and numerically similar can be justified from an empirical Bayes or robust Bayes perspective conditionally given S, and an estimator similar to $\tilde{\delta}_1$ is conditionally admissible given $\tilde{S}$. [See, e.g., Berger (1980, 1985).] However estimates qualitatively similar to $\tilde{\delta}_1$ cannot be unconditionally admissible. [See Remark 3.3.6 and Brown (1987)]. Thus, conditional use of (objectively or subjectively specified) formal Bayes estimators or empirical or robust Bayes methods may lead to inconsistency in the form of unconditional admissibility. It seems to me the conclusion is that none of these general paradigms should be applied conditionally without also taking into account the unconditional, frequentist structure of the situation.

Acknowledgments. The inadmissibility of δ_0 was presented (with a rather awkward proof involving an estimator somewhat like $\tilde{\beta}_3$) as part of my third Wald Lecture in August 1985. I want to thank C. Antoniak for a conversation there which helped me find the mathematical relation contained in Lemma 3.3.1 between this result and multivariate Stein estimation. I am also very indebted to L. Gleser for subsequently pointing out the significance and priority of Baranchik (1973) and for supplying the proof of Lemma 3.3.4. Some years ago R. Rubenstein introduced me to the control variables problem of Monte Carlo simulation, which for brevity can be though of as (3.3.1) in a different context. I am indebted to him for that introduction and for what I learned during our joint attempt to extend Rubenstein and Markus (1982).

REFERENCES

Baranchik, A. J. (1964). Multiple regression and estimation of the mean of a multivariate normal distribution. Technical Report No. 51, Dept. Statist., Stanford Univ.

Baranchik, A. J. (1973). Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. *Ann. Statist.* **1** 312–321.

Berger, J. O. (1975). Minimax estimation of location vectors for a wide class of densities. *Ann. Statist.* **3** 1318–1328.

Berger, J. O. (1976). Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Ann. Statist.* **4** 223–226.

Berger, J. O. (1980). A robust Bayes estimator and confidence region for a multivariate normal mean. *Ann. Statist.* **8** 545–571.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, Berlin.

Berger, J. O. and Srinivasan, C. (1978). Generalized Bayes estimators in multivariate problems. *Ann. Statist.* **6** 783–801.

Blyth, C. R. (1951). On minimax statistical decision procedures and their admissibility. *Ann. Math. Statist.* **22** 22–42.

Brown, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.* **37** 1087–1136.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903.

Brown, L. D. (1978). A contribution to Kiefer's theory of conditional confidence problems. *Ann. Statist.* **6** 59–71.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families.* IMS, Hayward, Calif.

Brown, L. D. (1987). An ancillarity paradox which appears in multiple linear regression (preliminary version). Technical Report, Cornell Statistics Center.

Cohen, A. (1965). Estimates of linear combinations of the parameters in the mean vector of a multivariate distribution. *Ann. Math. Statist.* **36** 78–87.

Copas, J. B. (1983). Regression, prediction, and shrinkage. *J. Roy. Statist. Soc. Ser. B* **45** 311–354.

Cox, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372.

Fox, M. (1981). An inadmissible best invariant estimator: the i.i.d. case. *Ann. Statist.* **9** 1127–1129.

Ghosh, M., Saleh, A. K. M. E. and Sen, P. K. (1987). Empirical Bayes subset estimation in regression models. Technical Report No. 281, Dept. Statist., Univ. of Florida.

Haff, L. (1979). An identity for the Wishart distribution with applications. *J. Multivariate Anal.* **9** 531–544.

Hodges, J. L. Jr., and Lehmann, E. L. (1951). Some applications of the Cramer–Rao inequality. *Proc. 2nd Berkeley Symp. Math. Statist. Prob.* 13–22.

Hudson, H. M. (1974). Empirical Bayes estimation. Technical Report No. 58, Dept. Statist., Stanford Univ.

Hwang, J. T. (1982). Certain bounds on the class of admissible estimators in continuous exponential families. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds) **2** 15–30. Academic, New York.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. 4th Berkeley Symp. Math. Statist. Prob.* **1** 311–319.

Judge, G. G. and Bock, M. E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometries.* North-Holland, Amsterdam.

Kiefer, J. (1976). Admissibility of conditional confidence procedures. *Ann. Statist.* **4** 836–865.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *J. Amer. Statist. Assoc.* **72** 789–827.

Narula, S. C. (1974). Predictive mean square error and stochastic regressor variables. *J. Roy. Statist. Soc. Ser. C* **23** 11–17.

OMAN, S. D. (1984). A different empirical Bayes interpretation of ridge and Stein estimators. *J. Roy. Statist. Soc. Ser. B* **46** 544–557.

PERNG, S. K. (1970). Inadmissibility of various "good" statistical procedures which are translation invariant. *Ann. Math. Statist.* **41** 1311–1321.

RUBINSTEIN, R. and MARKUS, R. (1982). *Improved estimation using control variables.* Report, Technion University, Haifa, Israel.

SAVAGE, L. J. (1976). On rereading R. A. Fisher (J. W. Pratt, ed.) *Ann. Statist.* **4** 441–500.

SCLOVE, S. L., MORRIS, C. and RADHAKRISHNAN, R. (1972). Nonoptimality of preliminary test estimators for the multinormal mean. *Ann. Math. Statist.* **43** 1481–1490.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. 3rd Berkeley Symp. Math. Statist. Prob.* **1** 197–206. Univ. California Press.

STEIN, C. (1959). The admissibility of Pitman's estimator of a single location parameter. *Ann. Math. Statist.* **30** 970–979.

STEIN, C. (1960). Multiple regression. In *Contributions to Probability and Statistics Essays in Honor of Harold Hotelling* (I. Olkin, et al., eds.) 424–443. Stanford Univ. Press.

STEIN, C. (1973). Estimation of the mean of a multivariate distribution. In *Proc. Prague Symp. on Asymptotic Statist.* 345–381. Charles Univ., Prague.

STRAWDERMAN, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42** 385–388.

TAKADA, Y. (1979). A family of minimax estimators in some multiple regression problems. *Ann. Statist.* **7** 1144–1147.

DEPARTMENT OF MATHEMATICS
WHITE HALL
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-7901

DISCUSSION

JAMES BERGER

Purdue University

The paper presents an exciting and rich mix of foundational issues concerning conditional reasoning and methodological developments involving improved estimation in multiple linear regression. My discussion will focus on the foundational issues, though certain features of the improved estimators will be used to illustrate some of the issues.

My first attempt at understanding the fundamental issue raised by the paper was along the following lines (sticking with the criterion of "admissibility" for preciseness):

> Ancillarity Paradox—A procedure which is conditionally admissible for each value of an ancillary statistic can be unconditionally inadmissible.

As I thought about it, however, this did not seem to capture the true novelty of the paper, because this ancillarity paradox has long been known, going back at least as far as the Cox example concerning testing with two randomly differing sample sizes. Brown notes that there is a difference between tests and estima-