This (3) is $(1 + r/c)\ \sigma^2/n$ when $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}$. The squared error loss (3) and distribution (2) for $\hat{\boldsymbol{\beta}}$ suggest choosing $\boldsymbol{\beta}^*$ to be Stein's estimator and leads to risk dominance and Brown's paradox.

With fixed $\overline{\mathbf{V}}$, and if $\boldsymbol{\beta}^*$ is Stein's estimator, the latter term in (1) is obtained from the component risk for Stein's rule. The risk (1) may be computed from component risk formulae of Baranchik (1964) and Efron and Morris (1972). It is

$$E(\alpha^* - \alpha)^2 = \frac{\sigma^2}{n} + \|\overline{\mathbf{V}}\|^2 \frac{\sigma^2}{n} R \tag{4}$$

and R is the component risk for Stein's rule

$$R = R_s + 2(r^2 - 4)\left(p - \frac{1}{r}\right) E \frac{2J}{(r + 2J)(r - 2 + 2J)}, \tag{5}$$

where J is Poisson with mean

$$\lambda = \frac{c\|\boldsymbol{\beta}\|^2}{2\sigma^2}, \qquad p = \frac{(\overline{\mathbf{V}}\boldsymbol{\beta})^2}{\|\overline{\mathbf{V}}\|^2\|\boldsymbol{\beta}\|^2}$$

and R_s is the average risk per component of Stein's estimator, having maximum of 1. The maximum of R in (4) and (5) is about $(r + 2)/4$, occurring when $\overline{\mathbf{V}}$ and $\boldsymbol{\beta}$ are collinear (when $p = 1$), and when 2λ is near r. Numerical values are in Efron and Morris (1972, Section 5).

If $p = 1/r$, (4) is given by Stein's risk. But for $p = 1$, the risk (5) starts above σ^2/n when $\boldsymbol{\beta} = \mathbf{0}$ and increases to a value exceeding (substantially, for large r) the risk of the MLE $\hat{\alpha}$ and then diminishes as $\boldsymbol{\beta} \to \infty$.

Perhaps this risk formula (5) will help us to understand the paradox. We should advocate these estimators for practical use only if we are sure they are appropriate. I do not think we know that yet.

REFERENCE

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

CENTER FOR STATISTICAL SCIENCES
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TEXAS
AUSTIN, TEXAS 78712

CHARLES STEIN

Stanford University

Both the mathematical results and the conceptual aspects of this paper are very interesting to me. Most of my remarks, which are numbered for convenient reference, are related to my rejection of the principle of conditionality, which is stronger than Brown's rejection of that principle.

1. Brown recognizes, immediately above Remark 3.4.1, that his estimator $\tilde{\delta}_3$ can have variance smaller than that of the maximum likelihood estimator δ_0 by an arbitrarily large factor when r is large compared with $n - r$ and the multiple correlation is small. It seems to me that this is an adequate reason for basing confidence intervals and tests on $\tilde{\delta}_3$ rather than on δ_0 if such intervals and tests are available. In practice there is some difficulty in studying the approximate distribution of $\tilde{\delta}_3$ unless both r and $n - r$ are large, but this does not affect the principle, and the computational difficulty should not be insuperable.
2. In trying to find an appropriate estimator of α it might be better to start with a Bayes estimator corresponding to a normal prior distribution with mean 0 for (α, β) with α independent of β and β having covariance matrix proportional to the identity, and with the variance of α approaching ∞. In order to obtain a usable estimator one might estimate the constant of proportionality in the prior variance of β (perhaps from the sample multiple correlation, for simplicity, if $n - r$ is not too small) or choose a slightly different prior distribution. This will again lead to computational difficulties. Intuitively it seems clear that the improvement over Brown's estimators would be large in some circumstances, especially if $n - r$ is small or negative.
3. Before taking a stand for (or even against) the principle of conditionality, for the version of the present problem referring to confidence intervals for α, it may be reasonable to examine in detail the conditional (and unconditional) behavior of the maximum likelihood estimator δ_0 and that of an alternative $\tilde{\delta}$ (which might be Brown's $\tilde{\delta}_3$). This study could be carried out mathematically or by simulation. For the present, a thought experiment must suffice.
4. Imagine that we have applied the two confidence interval procedures, say with probability 0.95, to a particular sample. What can we say about the resulting picture?

   (a) With probability greater than 0.9, the two intervals will be compatible in the sense that their intersection is nonempty. Thus the intervals based on $\tilde{\delta}$ will not ordinarily contradict those based on δ_0 but will claim greater precision. Of course the first sentence is an understatement. Roughly speaking, the intersection of the two intervals is a conservative 0.9 confidence interval and thus should be at least as large as a good 0.9 confidence interval on the average, though a bit more variable.

   (b) Nevertheless it is possible, though improbable, that the two intervals will be disjoint, even widely separated. In principle, in such a situation it is appropriate to use the theoretically superior procedure after checking the validity of the argument. However, in the present case, there is an additional argument in favor of intervals based on $\tilde{\delta}$ rather than δ_0.
5. As I understand it, the problem is to estimate the unconditional expectation α of the Y_i. We observe the V_i, because they are readily available, or because we recognize that they will help us estimate α. Since the conditional expectation of Y_i given V_i is assumed to be a constant plus a linear

function of V_i and the expectation of V_i is assumed to be 0, it follows that

$$\alpha = E^{V_i} Y_i - \beta' V_i \tag{1}$$

for some $\beta \in R^r$. If there is strong disagreement between confidence intervals based on δ_0 and those based on $\tilde{\delta}$, some doubt is cast on the assumptions leading to (1). Although both δ_0 and $\tilde{\delta}$ use (1), δ_0 uses it much more strongly. Also such disagreement is more probable (though still improbable under the assumptions) when $\tilde{\delta}$ is likely to be close to $\bar{Y}$, which is a valid estimate without the conditions on the distribution of V.

6. If the principle of conditionality is misguided, why do so many people find it appealing? I can think of a number of reasons:

   (a) In a small but conspicuous class of problems concerning confidence sets, it leads to reasonable answers in situations where unreasonable answers have been obtained from other principles.

   (b) In testing problems it arises from an attempt to attribute a significance level to the outcome of an experiment rather than to a test or a statistic.

   (c) It is sometimes justified on the basis of a verbal Bayesian argument.

   (d) Rejection of the principle of conditionality is sometimes confused with rejection of the use of conditional probability.

7. Consider the problem of obtaining confidence intervals for the location parameter of a one-dimensional uniform distribution or a Cauchy distribution with known scale. Confidence intervals based on the principle of conditionality coincide with those obtained from a uniform prior distribution and are reasonable. It is my impression that short unbiased confidence intervals in the sense of Neyman and those with minimum expected length seem unreasonable. Analogous results may hold for typical one-parameter problems, but they seem to fail, in varying degrees, for problems with more than one parameter, including problems with nuisance parameters.

8. In Cox's example for testing hypotheses, I think the difficulty arises from confusion of the ordinary meaning of the word "significance" with its use as a technical term in statistics. We can assign a level of significance to a test or even to the observed value of a statistic (a different but related notion), but not ordinarily to the outcome of an experiment. As indicated in 7 above, I would be inclined to use the confidence intervals obtained by the principle of conditionality in this example, but not the test.

9. The passage of Savage (1976) that Brown quotes in Section 5 is an example of a verbal Bayesian argument for the principle of conditionality. However, I think Savage's agreement with the principle of conditionality is not as strong as Brown suggests. Furthermore, in the present situation, the Bayesian who assumes α and β independent, with uniform prior for α and a reasonable orthogonally invariant prior distribution for β, and conditions on V (and Y) obtains qualitatively the same result as Brown does by not conditioning on V. Thus a serious Bayesian argument cannot support the non-Bayesian interpretation of the principle of conditionality.

10. I regret that I have not had time to do the mathematical work that would, I believe, support some of the above statements.
11. It should also be remembered that the literature on the principle of conditionality is extensive.
12. A general principle, like a mathematical assertion beginning with a universal quantifier, can be refuted by a single counterexample but cannot be validated or proved by any number of special examples.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

REJOINDER[1]

LAWRENCE D. BROWN

Explanations, etc. Several discussants have offered supplementary explanations for the inadmissibility result of Section 3.3 (*Casella*, *Copas*, *Efron*, *Gleser*, *Morris*). Each of the explanations is somewhat different and each adds further understanding.

Gleser goes further and gives a useful extension of my results in the style of my Lemmas 3.3.1 and 3.3.3. Consider the situation discussed in my Section 4.2 where it is desired to estimate the linear function $\kappa = a\alpha + b'\beta$ in the regression setting. Then, if $r \geq 3$, *Gleser*'s Theorem 1 can be applied via his formula (5) to yield a specific, useful estimator dominating $\delta_0 = a\hat{\alpha} + b'\hat{\beta}$. The existence of a dominating estimator was already established in my Theorem 4.2, even for $r = 2$, but no usable formula was given.

Lu demonstrates that the general nature of the inadmissibility phenomena here is not significantly dependent on the form of the loss function. Insofar as his results for L_c are not directly constructive (analogous to my Theorems 2.2.1 and 3.2.2 for squared error) they point to the important question of constructing estimators in the regression setting which usefully dominate δ_0 under L_c.

Limited translation estimators. *Morris* (explicitly) and *Efron* (implicitly) each raise the issue of modifying the proposed estimators to limit maximum coordinatewise risk. (This appears to be the joint occurrence of conditionally independent but marginally highly correlated events!) *Berger* also makes this suggestion. This seems reasonable, particularly in view of the numerical results *Berger* mentions. However it is important to understand the justification for this suggestion before putting it into practice.

To do so consider the usual multiple normal means estimation problem and the positive-part James–Stein estimator, which is given by d^*_+ of (2.1.7) for $\Sigma = \Omega = I$ and $\rho = p - 2$. For moderate $p \geq 3$ this is known to approximate a

[1]Research supported in part by NSF DMS 88-09016.