REFERENCES

BASU, D. (1964). Recovery of ancillary information. In *Contributions to Statistics* 7–29. Pergamon Press, Oxford. Republished in *Sankhyā* **26** 3–16 (1964). Also in *Statistical Information and Likelihood, A Collection of Critical Essays by Dr. D. Basu* (J. K. Ghosh, ed.). Springer, New York (1988).

BASU, D. (1981). On ancillary statistics, pivotal quantities, and confidence statements. In *Topics in Applied Statistics* (Y. P. Chaubey and T. D. Disivedi, eds.) 1–20. Concordia University, Montreal. Also in *Statistical Information and Likelihood, A Collection of Critical Essays by Dr. D. Basu* (J. K. Ghosh, ed.). Springer, New York (1988).

BUEHLER, R. (1982). Some ancillary statistics and their properties (with discussion). *J. Amer. Statist Assoc.* **77** 581–593.

FISHER, R. A. (1936). Uncertain inference. *Proc. Amer. Acad. Arts Sci.* **71** 245–258.

ROBINSON, G. K. (1979). Condition properties of statistical procedures. *Ann. Statist.* **7** 742–755.

BIOMETRICS UNIT
337 WARREN HALL
CORNELL UNIVERSITY
ITHACA, NY 14853

J. B. COPAS

University of Birmingham

The paper gives a counterexample against two common assumptions about shrinkage estimation: First, that shrinkage applies only to vector parameters with a single loss function allowing exchangeability of errors between component estimates; second, that shrinkage estimates are necessarily biased. In this important paper Dr. Brown provides a shrinkage estimate which is both scalar and unbiased. Particularly interesting and surprising is the finding that improved estimation of α is possible only when the mean of the V_i's is known. Are there intuitive grounds for expecting this to be so?

In Section 3 of the paper, suppose that the values of the V_i's are temporarily lost. Can anything then be said about α? Usually the answer is no, but if we have the additional information that the V_i's are *known* to have mean zero, then $\hat{\alpha}_0 = \bar{Y}$ becomes a (globally) unbiased estimate. With the full data $\hat{\alpha}$ is available as a second unbiased estimate, suggesting that we could do even better with the combined unbiased estimate

$$\hat{\alpha}(\lambda_1) = (1 - \lambda_1)\hat{\alpha}_0 + \lambda_1 \hat{\alpha}.$$

Evaluating global moments (averaging over Y as well as V) we obtain the variances of $\hat{\alpha}_0$ and $\hat{\alpha}$ as $n^{-1}(\sigma^2 + \beta'\beta)$ and $n^{-1}\sigma^2(1 + r/(n - r - 2))$, respectively, with covariance $n^{-1}\sigma^2$. The variance of $\hat{\alpha}(\lambda_1)$ is therefore minimized when

$$\lambda_1 = \frac{\beta'\beta(n - r - 2)}{\beta'\beta(n - r - 2) + r\sigma^2}.$$

A (globally) unbiased estimate of $\beta'\beta$ is

$$\hat{\beta}'\hat{\beta} - \frac{r\sigma^2}{n - r - 2}.$$

Constructing the corresponding estimate of λ_1 then suggests the estimate

$$\hat{\alpha}_1 = \hat{\alpha}(\hat{\lambda}_1) = \overline{Y} - \left(1 - \frac{r\sigma^2}{(n - r - 2)\hat{\beta}'\hat{\beta}}\right)\overline{V}\hat{\beta}.$$

This is very similar to Dr. Brown's estimate using $\tilde{\beta}_2$ in (3.3.4). Note that the argument cannot apply if the mean of the V_i's is unknown.

A second argument is based on prediction. This makes a different, but related, assumption, namely that a future value of v can be assumed to arise with the same mean and variance as the sample mean and variance of the V_i's in the data. The idea is that validating a predictor in terms of its performance on such a new observation is similar in spirit to cross-validation on the observations in the sample. Adding normality, suppose that $v \sim N(\overline{V}, n^{-1}S)$ and that a new observation y is then sampled from $N(\alpha + v\beta, \sigma^2)$. The least squares predictor of y is $\hat{y} = \hat{\alpha} + v\hat{\beta}$. Then conditional on Y and V, $(y, \hat{y})$ is bivariate normal from which we obtain

$$E(y|\hat{y}) = \alpha + \overline{V}\beta + \lambda_2(\hat{y} - \overline{Y}),$$

where $\lambda_2 = \hat{\beta}'S\beta/\hat{\beta}'S\hat{\beta}$. Now conditional on V we have

$$E(\hat{\beta}'S\hat{\beta}) = \beta S\beta + r\sigma^2$$

and

$$E(\hat{\beta}'S\beta) = \beta S\beta.$$

Thus the natural estimate of λ_2 is $\hat{\lambda}_2 = 1 - r\sigma^2/\hat{\beta}'S\hat{\beta}$, which with α and β estimated in the obvious way suggests that the above conditional expectation is estimated by

$$\tilde{y}(v) = \overline{Y} + \hat{\lambda}_2(v - \overline{V})\hat{\beta}.$$

Now the value of α is the expected value of y at $v = 0$; hence the estimate

$$\hat{\alpha}_2 = \tilde{y}(0) = \overline{Y} - \left(1 - \frac{r\sigma^2}{\hat{\beta}'S\hat{\beta}}\right)\overline{V}\hat{\beta},$$

similar to Dr. Brown's estimate using $\tilde{\beta}_3$ in (3.3.5). Note that if no assumption can be made about the process generating the V_i's, then the above argument based on the similarity of present and future values of v would be unreasonable, and hence the motivation for $\hat{\alpha}_2$ would not apply.

Of course, these arguments prove nothing about admissibility but do suggest that the necessity for the known mean of the V_i's is not unreasonable.

SCHOOL OF MATHEMATICS AND STATISTICS
THE UNIVERSITY OF BIRMINGHAM
P.O. BOX 363
BIRMINGHAM B15 2TT
UNITED KINGDOM

B. EFRON

Stanford University

Here is a slightly simpler version of Brown's nice paradox: the statistician observes $X \sim N_p(\mu, I)$, $p \geq 3$, and also an integer J that equals $j = 1, 2, 3, \dots, p$ with probability $1/p$, independently of X. It is desired to estimate μ_j with squared-error loss. Then J is ancillary, and conditional on $J = j$ the obvious estimate $d_0(X, j) = X_j$ is admissible and minimax. Unconditionally, however, the Jth coordinate of the James–Stein estimate,

$$d_1(X, J) = \left[1 - (p - 2)/\|X\|^2\right] X_J,$$

dominates $d_0(X, J)$, with $E[d_1(X, J) - \mu_J]^2 < E[d_0(X, J) - \mu_J]^2$ for all vectors μ.

In other words, Brown has restated Stein's paradox, that d_1 dominates d_0 in terms of total squared error loss, in an interesting way that casts some doubt on the ancillarity principle.

[The example above does not look much like Brown's regression paradox, but we can fix things up by supposing that given $J = j$ the statistician also observes $X_0 \sim N(\alpha + \mu_j, 1)$, independent of $X \sim N_p(\mu, I)$, the goal now being to estimate α with squared-error loss. Then $\hat{\alpha}_1 = X_0 - d_1(X, J)$ dominates $\hat{\alpha}_0 = X_0 - d_0(X, J)$ unconditionally but not conditionally. This situation might arise if X_j was the placebo response of patient j on some physiological scale and X_0 was patient j's response when given a treatment of interest; we placebo-test p patients and then choose one at random to receive the treatment.]

Why do we intuitively accept the ancillarity principle in Cox's example, Section 5, but doubt it in the example above, or in Brown's regression paradoxes? I believe that the answer has more to do with single versus multiple inference than with hypothesis testing versus estimation.

Notice that $d_0(X, j)$ disregards all of the data except X_j. There is nothing in the ancillarity principle to justify this. All that ancillarity says is that we should do our probability calculations conditional on $J = j$. In Cox's example on the other hand, the conditional solution makes use of all the data and the ancillarity principle works fine.

Even when the choice $J = j$ is totally nonrandom it is not obvious that d_0 is preferable to d_1. The real question is whether or not the ensemble estimation gains offered by d_1 are relevant to the specific problem of estimating μ_j.