HUBER, P. J. (1983). Minimax aspects of bounded-influence regression. *J. Amer. Statist. Assoc.* **78** 66–80.

JOBSON, J. D. and FULLER, W. A. (1980). Least squares estimation when the covariance matrix and parameter vector are functionally related. *J. Amer. Statist. Assoc.* **75** 176–181.

KRASKER, W. S. and WELSCH, R. E. (1982). Efficient bounded-influence regression estimation. *J. Amer. Statist. Assoc.* **77** 595–604.

DEPARTMENT OF STATISTICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27514

B. EFRON

Stanford University

The remarks that follow are mainly critical, but that is not unusual when statisticians discuss difficult new areas of research. My criticism is not meant to obscure the paper's many positive achievements: the neat development of resampling methods for the linear regression problem, in particular Theorem 2; the extended class of weighted jackknives introduced in Section 4, and their justification in Theorem 3; and the intriguing suggestion in Section 8 for a more general weighted jackknife based on the Fisher information. The paper's main fault, in my opinion, is not the absence of interesting new ideas but rather an overinterpretation of results, which leads to bold distinctions not based on genuine differences.

(A) I reran part of the simulation experiment of Section 10, exactly as described except for the following change: Instead of taking the $e_i \sim N(0, x_i/2)$, I took them $N(0, |x_i - 5.5|)$. This gives nearly the same set of variances for the errors e_i, but with the large variances occurring at both ends of the x range, rather than just at the right end. Only the estimation of $\mathrm{Var}(\beta_0)$ (actually equal 3.64 in this situation) was considered, and only by the two estimators $v_{J(1)}$, definition (5.1), and $\hat{v}$, definition (2.9).

Here are summary statistics for 400 Monte Carlo trials:

	mean	st. dev.	rms
$v_{J(1)}$	3.47	3.14	3.14
$\hat{v}$	2.40	1.20	1.73

(rms indicates root mean square error). Now $\hat{v}$, the ordinary estimator (and also the "residual bootstrap" estimator v_b (2.9)), is biased sharply downward instead of upward as in Table 1; $v_{J(1)}$ is nearly unbiased, as it was designed to be.

However $v_{J(1)}$ *is much more variable than* $\hat{v}$, having nearly three times the standard deviation and twice the rms error for estimating $\mathrm{Var}(\beta_0)$. The percentiles of the two Monte Carlo distributions

	5%	10%	16%	50%	(true)	84%	90%	95%
$v_{J(1)}$	0.57	0.83	1.02	2.47	(3.64)	6.15	7.80	9.63
$\hat{v}$	0.88	1.12	1.27	2.14	(3.64)	3.65	4.06	4.56

show that $v_{J(1)}$ is genuinely bad here, even compared to $\hat{v}$, which was never designed to handle the situation at hand. Usually it is not worthwhile to push too hard for unbiasedness, particularly in estimating variances, where the most important application, approximate confidence intervals, refers to standard deviation rather than variance anyway.

Of course this one simulation does not prove that $v_{J(1)}$ is bad in general. The same cautionary remark applies to the results in Section 10! [Theoretical considerations (see remark G below) indicate that $v_{J(1)}$ will be more variable than $\hat{v}$ in most cases. When the σ_i^2 are all equal, for example, the rms error of $v_{J(1)}$ exceeds $\hat{v}$ by 41%, for the estimation of $\mathrm{Var}(\beta_0)$ in the quadratic regression of Section 10.]

(B) The jackknife and bootstrap are general-purpose devices, not specifically adapted to take advantage of a special model like (2.1). Comparisons with specially adapted methods such as the author proposes are misleading if this is not made clear. For example, the confidence interval method PBOOT in Table 3 does not incorporate a student's t-correction, as do the entries above it in the table. If a t-correction is applied to PBOOT, say widening interval (2.10) by a factor of $t_9^{(0.025)}/z^{(0.025)} = 1.18$ about its central point, then the "equal variance" entries for PBOOT in Table 3 will nearly equal the nominal value 0.95.

Why not always widen PBOOT by a student's t factor? Because in general problems, as opposed to the special case of linear regression with normally distributed errors, it is not clear how to choose such a factor. General-purpose nonparametric confidence intervals are discussed in Section 7 of Efron and Tibshirani (1986). There has been substantial progress in this area, made by many authors using different techniques, but the problem is far from solved.

(C) Attaining the nominal coverage probability does not make a proposed confidence interval correct. Notice in Table 4 that the intervals for VCJ8 are *much* too long (as the author points out), and could not be used in practice. The six methods in the simulation study that start with "V" are symmetrically centered at the point estimate $\hat{\theta}$. As discussed in Efron and Tibshirani (1986), the asymmetry of genuinely correct intervals, such as Fieller's construction, is an order of magnitude larger effect than the student's t-correction.

(D) The paper treats the various methods as competitors, but in fact their similarities are more striking than their differences. Here is a schematic diagram of bootstrap methods in general, Efron and Tibshirani (1986), that helps relate the estimators:

$$\begin{array}{ccccc} P & \to & y & \Rightarrow & \hat{P} \to y^* \\ & & \downarrow & & \downarrow \\ & & \hat{\beta} & & \hat{\beta}^* \end{array}$$

The observed data y comes from a specified but unknown probability mechanism P, in this case the linear model $y = X\beta + e$ (2.1), with unknown parame-

ters β and $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. We want to assess the variability of a statistic computed from y, here $\hat{\beta} = (X^TX)^{-1}X^Ty$.

Bootstrap methods proceed by first estimating the entire probability mechanism P, say by $\hat{P} = (\hat{\beta}, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_n^2)$ in this case; then by resampling data vectors y^* from $\hat{P}$; recalculating the statistic of interest, $\hat{\beta}^* = (X^TX)^{-1}X^Ty^*$; and finally using the observed variability of the $\hat{\beta}^*$ values to estimate the variability of $\hat{\beta}$. In situation (2.1) the bootstrap calculations can be carried out theoretically, without recourse to Monte Carlo, giving the bootstrap variance estimate

$$v_{\text{BOOT}} = (X^TX)^{-1}\sum_1^n \hat{\sigma}_i^2 x_i x_i' (X^TX)^{-1}$$

for $\hat{\beta}$.

The estimate $\hat{v} = v_b$ (2.9), uses $\hat{\sigma}_i^2 = \hat{\sigma}^2$. The estimate $v_{J(1)}$ (5.1) uses $\hat{\sigma}_i^2 = r_i^2/(1 - w_i)$. The Hinkley estimate $v_{H(1)}$ (2.6) (which, incidentally, performed somewhat better than $v_{J(1)}$ in the simulation experiment of remark A) uses $\hat{\sigma}_i^2 = r_i^2/(1 - k/n)$. Professor Wu suggests another variation in Section 7. A great variety of other possibilities is possible, for example using $\hat{\sigma}_i^2 = cr_i^2 + (1 - c)\hat{\sigma}^2$, where c is a Stein-like shrinkage factor.

(E) The discussion beginning at (6.14) of the "unweighted bootstrap," the bootstrap which resamples pairs (y_i, x_i) rather than residuals, is misleading. If the parameter of interest μ is the intercept of the true regression line at the particular point $x = \bar{x}$, then μ^* is *not* $\bar{y}^* = \sum_1^n y_i^*/n$, but rather $\tilde{\mu}$ as described following (6.17). The unweighted bootstrap estimate of variance is quite close to the usual answer (6.15) in this case.

(F) Notice that the usual jackknife estimate v_J and the unweighted bootstrap estimate v^* have nearly identical biases in Table 1. Theorem 6.1 of Efron (1982), which justifies the jackknife as a linear approximation to the bootstrap, is quantitatively accurate in this case. The large biases observed for entry (1, 1), the estimated variance of the intercept $\hat{\beta}_0$, occurs for a simple reason: In resampling the pairs (y_i, x_i), the unweighted bootstrap uses data sets with varying collections of x-values, whereas $\text{Var}(\hat{\beta})$ in Table 1 refers to the 12 fixed values of x given at the beginning of Section 10.

In other words, v_J and v^* estimate unconditional rather than conditional variances. This is a higher order effect that ordinarily does not seem to be very important. Here it is important because of the small degrees of freedom (9) and the fact that β_0 is the intercept at 0, outside the range of support of the x_i, so that $\hat{\beta}_0$ is an extrapolation. For situation (2.1) it is easy enough to run the bootstrap conditional on the x values, as in remark D. In other situations (see for example Reid (1981)), conditional bootstrapping is not at all obvious.

(G) The problem considered in this paper, estimating the variance of a function of $\hat{\beta}$ in the heteroscedastic situation (2.7), is interesting in its own right, independent of jackknife/bootstrap considerations. The estimators $v_{J(1)}, v_{H(1)}$

and $\hat{v}$ are of the form $\sum_{i=1}^{n} c_i r_i^2$, where the constants c_i depend on X but not on the σ_i^2. For the situation in Section 10, the vectors $\mathbf{c} = (c_1, c_2, \ldots, c_{12})$ are

$v_{J(1)}$: (0.67, 0.28, 0.12, 0.04, 0.00, 0.00, 0.01, 0.06, 0.08, 0.06, 0.02, 0.46)

$v_{H(1)}$: (0.52, 0.28, 0.13, 0.04, 0.01, 0.00, 0.02, 0.09, 0.09, 0.06, 0.02, 0.12)

$\hat{v}$: (0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11, 0.11)

This raises several interesting questions, some of which are considered in the MINQUE literature:

(i) For a given vector of variances $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$, what is the best choice of the vector $\mathbf{c}$, say to minimize rms error?

(ii) Given a set of possible variance vectors, is there a preferred general choice of $\mathbf{c}$?

(iii) Is there an adaptive way of selecting $\mathbf{c}$ from the observed data, as suggested at the end of remark D?

(iv) Wu's estimators $v_{J,r}$, $r > 1$, involve quadratic forms $\sum_i \sum_j c_{ij} r_i r_j$. Is there any real advantage to using the cross-terms $r_i r_j$, or does this just add to the variability of the estimator?

REFERENCES

EFRON, B. and TIBSHIRANI, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other methods of statistical accuracy (with discussion). *Statist. Sci.* **1** 54–77.

EFRON, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM, Philadelphia.

REID, N. (1981). Estimating the median survival time. *Biometrika* **68** 601–609.

DEPARTMENT OF STATISTICS
SEQUOIA HALL
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

JOSEPH FELSENSTEIN

University of Washington

The emphasis the author places on using nonstandard subset sizes in jackknife procedures is important for at least one other reason. In some applications of resampling methods we are estimating an entity that lives in a space in which extrapolation is essentially impossible. Wu's equation (4.4) takes the estimate obtained from analysis of a sample and extrapolates its deviation from the overall estimate. However, if the space does not admit of extrapolation, then a choice of a subset size of $(n + k - 1)/2$ eliminates the extrapolation entirely. The value of $\tilde{\beta}_s$ is then the same as $\hat{\beta}_s$. By resampling and collecting a set of values of $\hat{\beta}_s$ we get variation that we take to be typical of the sampling variation in the estimate.

The example that brings this to mind is placing confidence intervals on phylogenies (evolutionary trees), to which I have applied a bootstrap technique