SRIVASTAVA, M. S. and CHAN, Y. M. (1985). A comparison of bootstrap method and Edgeworth expansion in approximating the distribution of sample variance—one sample and two sample case. Technical report 23, Univ. of Toronto.

SUBRAHMANYAM, M. (1972). A property of simple least squares estimates. *Sankhyā Ser. B* **34** 355–356.

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO M5S 1A1
CANADA

ROBERT TIBSHIRANI[1]

University of Toronto

Professor Wu has made a substantial contribution to a difficult area: the study of resampling methods in regression. The idea of weighted jackknife and bootstrap estimates of variance is an intriguing and potentially useful one. However, I feel that this paper falls short of providing any definitive answers because it overemphasizes unbiasedness and fails to address some important statistical issues. I will elaborate on these points as they relate to estimates of variance in regression, then I will conclude with a few remarks about confidence procedures. Despite the mostly critical comments that follow, I want to make it clear that I wholeheartedly endorse one of the major thrusts of the paper, namely Professor Wu's recommendation that "important features of a problem should be taken into account in the choice of resampling methods." This is good advice—it is just not clear yet how to do this in many problems.

Before computing an estimate of variance in a regression, there are two important questions that we should ask: (1) is our model adequate for the data and (2) do we want an estimate of the conditional or unconditional variance? Let us consider the first point. Given that we are going to use a linear model, the two main types of model inadequacy are misspecification of the mean of the response and nonhomogeneity of errors. Professor Wu assumes throughout that the mean part of the model is correctly specified. In fact, it is when the mean is misspecified that the unweighted procedures can still give a reliable estimate of variance. This is what I believe Efron and Gong meant in their claim about the robustness of the unweighted bootstrap. We will return to this point later, but for now we will assume that the mean is specified correctly, with possible heterogeneity of error variance.

Regarding the second point, Professor Wu uses the *conditional* variance, that is, the variance conditional on the observed X's, as his gold standard. An alternative gold standard is the unconditional variance, averaging over the marginal distribution of the X's. Which is the "correct" variance is an arguable point when the X's are not fixed by design, although ancillarity arguments can

[1] Research supported by the Natural Sciences and Engineering Researc: Council of Canada.

be given for the conditional variance in the normal case. Now the unweighted jackknife and bootstrap procedures are estimates of the unconditional variance: the bootstrap simply replaces the unknown joint distribution of the X's and Y by their empirical joint distribution, and the jackknife is a linear approximation to the bootstrap (Efron (1979)). Not surprisingly, the estimators of the unconditional variance overestimate the conditional variance, because they include the marginal variation of the X's. Professor Wu's clever idea is to weight the jackknife and bootstrap by the determinant of the subsample design matrix so that they approximate the conditional variance estimate. In particular he shows that his weighted jackknife is unbiased for the true conditional variance. Note however that the difference between the unconditional and conditional variances is not usually large (although it is in the small sample example given in Section 10). In fact if we look closely at Professor Wu's analysis of Section 5, where he shows that $E(v_{J(1)}) = \operatorname{var}(\hat{\beta})(1 + O(1/n))$, we find that this is also true for the (unconditional) estimator v_J, and hence conditioning affects the bias only by $O(1/n)$. To see this, note the formula for v_J is the same as (5.2) except that $r_i^2/(1 - w_i)$ is replaced by $r_i^2/(1 - w_i)^2$. Now w_i is assumed to be $O(1/n)$ and thus

$$E(r_i^2) = (1 - w_i)\sigma_i^2 + O(1/n) = (1 - w_i)^2\sigma_i^2 + O(1/n)$$

and the result follows.

It is also important to remember that unbiasedness of an estimator is not everything: variance is also important. For example, if the variance estimate is to be used for form a confidence interval for a regression parameter, then the accuracy of the resultant confidence interval will be a function of the mean squared error of the corresponding estimate of standard error. To pursue this point, I reran part of Professor Wu's first simulation study. The regression coefficients β were chosen to be (1, 1, 5). (Note that the bias of the estimates is independent of β but the mean squared error is not.) The root mean squared error of the estimates is shown in Table 1. The unweighted and weighted bootstrap estimates were not included: they are likely to perform similarly to the unweighted and weighted jackknife estimates, respectively. In all cases, the biases of the estimators (not shown) were close to those obtained by Professor Wu. In both problems $v_{J(1)}$ easily outperforms the unconditional estimate v_j but has higher mean squared error than either v or $v_{H(1)}$, especially in the equal variance case.

It is interesting to take a closer look at the unequal variance situation. The estimators v, v_J, $v_{J(1)}$ and $v_{H(1)}$ are all of the form $(X^tX)^{-1}\Sigma_1^n\hat{\sigma}_i^2 x_i x_i^t (X^tX)^{-1}$ for some estimates $\hat{\sigma}_i^2$. They use, respectively, $\hat{\sigma}^2$, $r_i^2/(1 - w_i)^2$, $r_i^2/(1 - w_i)$ and $r_i^2/(1 - k/n)$ for $\hat{\sigma}_i^2$. Professor Wu shows that $v_{J(1)}$ has bias only $O(1/n)$, but it appears from Table 1 that variance is the dominating factor. Both Hinkley's estimator, which replaces the w_i in $v_{J(1)}$ by its average value, and v which goes further by replacing $r_i^2/(1 - k/n)$ by its average value, outperform $v_{J(1)}$. Ironically, none of the estimators were able to improve substantially on v, which assumes equal error variances. As an alternative, I tried using the scaled cross-validated residual $r_{(i)}^2/(1 + w_i)$, as an estimate of σ_i^2. ($r_{(i)}$ is y_i minus the fit

TABLE 1
Variance of the least squares estimator and root mean squared error of variance estimators.

	(1, 1)	(1, 2)	(1, 3)	(2, 2)	(2, 3)	(3, 3)
			Equal variances			
var($\hat{\beta}$)	1.01	−0.42	0.03	0.21	−0.02	0.00
rmse(v)	0.48	0.20	0.00	0.10	0.00	0.00
rmse(v_J)	1.62	0.74	0.07	0.40	0.04	0.00
rmse($v_{J(1)}$)	0.83	0.31	0.03	0.13	0.01	0.00
rmse($v_{H(1)}$)	0.73	0.27	0.02	0.11	0.01	0.00
			Unequal variances			
var($\hat{\beta}$)	1.50	−0.79	0.08	0.48	−0.05	0.01
rmse(v)	1.28	0.48	0.04	0.24	0.02	0.00
rmse(v_J)	3.45	2.10	0.24	1.32	0.16	0.02
rmse($v_{J(1)}$)	1.00	0.52	0.05	0.31	0.03	0.00
rmse($v_{H(1)}$)	0.78	0.41	0.04	0.25	0.03	0.00

at x_i with the ith point removed, and can be shown to equal $r_i/(1 - w_i)$.) There is reason to believe that this should be a better estimate of σ_i^2 than $r_i^2/(1 - w_i)$. However, it performed much worse than $v_{J(1)}$ in the problem of Table 1, displaying large variability, even if the $\hat{\sigma}_i^2$'s were smoothed.

Table 2 shows the results of a number of variance estimators, with the same true model as before, but with the quadratic term left out of the fitted models. Also, instead of fixing a set of X's, the data were generated by sampling the X's with replacement from the original set of X values, then adding a $N(0, 1)$ random variable to the quadratic curve. The table shows the median bias and median absolute deviation of each estimate from the true *unconditional* variance. (Medians were used because both jackknife methods occasionally produced a very large variance for the (1, 1) entry.) We see that v_J has much less bias than $v_{J(1)}$ or v; this is to be expected because v_J is a linear approximation to the unweighted bootstrap, which makes no special use of the form or correctness of the regression model. However, v_J shows greater variability than the other estimators and hence the median absolute deviations are all comparable. (I believe that in larger samples v_J would have smaller variability and be preferable

TABLE 2
Analysis of incorrect model case.

	(1, 1)	(1, 2)	(2, 2)
var($\hat{\beta}$) (unconditional)	635.02	−153.27	43.70
median bias (v)	−279.00	96.04	−30.98
median bias (v_J)	93.54	−8.98	−7.50
median bias ($v_{J(1)}$)	−160.78	47.14	−18.00
MAD (v)	278.99	96.04	30.98
MAD (v_J)	366.45	83.01	24.97
MAD ($v_{J(1)}$)	303.86	76.87	24.02

to v or $v_{J(1)}$.) Note that all of the estimators considered here are poor estimators of true *conditional* variance, in particular $\hat{\sigma}^2$ averaged about 60, which is 60 times too large.

We can summarize these results as follows (being careful not to put too much faith in the results of one simulation study). If the mean part of the model is correctly specified, then v, $v_{J(1)}$ or $v_{H(1)}$ provide reasonable estimates of the conditional variance. Neither $v_{J(1)}$ or $v_{H(1)}$ improve upon the standard estimator v even under heteroscedasticity. The (unweighted) estimators v_J (and by implication v^*) are poor estimators of the conditional variance if large conditioning effects are present. If the mean is incorrectly specified, then none of the estimators are good estimators of the conditional variance. The conditional estimates are very biased for the unconditional variance, while the unconditional estimate v_J is not very biased but shows large variability.

I conclude with a few remarks:

1. Of the seven confidence interval procedures discussed in the second Monte Carlo study, only the jackknife and bootstrap percentile methods can produce intervals that are asymmetric about $\hat{\theta}$. This can be important because good small small confidence intervals are often asymmetric.
2. Professor Wu has reported the coverage and length of the intervals but not how closely the endpoints of each matches the Fieller interval endpoints. Thus we can not properly evaluate their performance. For example, an interval could be much shorter than the Fieller interval on the left and longer by the same amount on the right. The length and overall coverage would be just right but the interval would be a poor one.
3. In remark 4 of Section 11, Professor Wu says that refinements of the bootstrap and jackknife histogram are called for, because of their poor performance in the simulation study. If indeed they performed poorly, I think that it is due to the fact that they are unconditional methods (in the sense described earlier) that are being assessed by their conditional coverage. The refinements of Efron (1985, 1986) are improvements on the percentile interval to achieve second-order correctness. However, Efron (1985) has shown that in the simple (nonregression) version of the Fieller–Creasy problem, the percentile interval is second-order correct, so the refinements are unlikely to help here.
4. The warning in Section 6.2, about careless use of the unweighted bootstrap, is valuable. In the example given, the parameter of interest is $\mu = \alpha + \beta\bar{x}$ and thus depends in an explicit way on the observed set of X's. It makes no sense to use an unconditional method like the unweighted bootstrap or jackknife for such a parameter. Focussing instead on α (which is defined independently of the observed X's) alleviates the problem. In general, we can heed the following warning: When using a resampling method, we must think carefully about what components of the problem are to be considered fixed. The resampling method must take this into account.
5. The general problem of resampling residuals (Section 7) can be expressed in another way. As Professor Wu notes, bootstrapping the residuals doesn't work

under heteroscedasticity because it implicitly assumes that the true residuals all have the same distribution. We can assume instead that the distributions are different, say H_i for the ith residual and estimate each H_i in some way. Professor Wu's suggestion amounts to estimating the joint distribution of the residuals by a distribution having marginal mean 0 and variance $r_i^2/(1 - w_i)$. There are two problems with his suggestion: a) The resampled residuals are uncorrelated but not necessarily independent, as they should be, and more importantly, b) only the first two moments of this distribution are specified, so there is no hope of capturing higher-order effects. A method that estimated each H_i with the empirical distribution function of the residuals in some neighborhood of the ith point might hold more promise.

Acknowledgment. I would like to thank Brad Efron for suggesting mean squared error as a relevant measure.

REFERENCES

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* **72** 45–58.

EFRON, B. (1986). Better bootstrap confidence intervals. To appear in *J. Amer. Statist. Assoc.*

DEPARTMENT OF PREVENTIVE MEDICINE
AND BIOSTATISTICS AND DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO M5S 1A8
CANADA

NEVILLE WEBER

University of Sydney

Wu's paper should be praised for clarifying the relationship between the various weighted and unweighted versions of the jackknife and the bootstrap and for giving simple examples to demonstrate the shortcomings of various methods. The paper also stresses the role of the jackknife in regression analysis as a means of obtaining an estimator for the covariance matrix of the least squares estimator, $\tilde{\beta}$, which is robust against heteroscedastic errors. This feature of the jackknife has not received enough emphasis in regression literature.

As noted by Weber (1986), Hinkley's weighted jackknife variance estimator, $V_{H(1)}$, is effectively the robust estimator proposed by White (1980), commonly used in econometrics. The new estimator $V_{J(1)}$ has the consistency property of $V_{H(1)}$, but also has the advantage of being unbiased when the model has homoscedastic errors.

The bootstrap procedure in regression does not lead to a robust, consistent estimator of the covariance matrix of $\hat{\beta}$. The bootstrap method based on resampling the residuals has the obvious shortcoming of imposing a linear model structure on the resampled values, forcing the error terms to be independent and identically distributed. Such a procedure loses any variation in the distributions