M. S. SRIVASTAVA

University of Toronto

This is technically an impressive article that is full of innovative ideas on resampling methods. Although the unbalanced nature of the regression model was first recognized by Hinkley (1977), it is this paper that brings to light various pitfalls in the estimation of the variance of regression estimates by different resampling schemes. Motivated by a representation for the least-squares estimator, first given by Subrahmanyam (1972), although not as rigorous as in this paper, Wu provides robust estimates of the variance. It also shows the failure of the bootstrap method.

The failure of the bootstrap method is not new. Many examples exist in the literature, even when the observations are not as unbalanced as in the regression case (see, for example, Singh (1981)). Another example is when the vectors (X_i, Y_i) are independently distributed with mean vector zero, $\mathrm{Var}(X_i) = \sigma_{11}$, $\mathrm{Cov}(X_i, Y_i) = \sigma_{12}$ and $\mathrm{Var}(Y_i) = a_i\sigma_{22}$, a_i's known, $i = 1, \dots, n$. An estimate of σ_{22} is given by

$$\hat{\sigma}_{22} = n(n-1)^{-1}(\Sigma a_i)^{-1}\Sigma(Y_i - \overline{Y})^2,$$

with a bootstrap estimate

$$\hat{\sigma}_{22}^* = n(n-1)^{-1}(\Sigma a_i)^{-1}\Sigma(Y_i^* - \overline{Y}^*)^2.$$

The bootstrap estimate of $\mathrm{Var}(\hat{\sigma}_{22})$ is not consistent, unless $a_i \to 1$.

It is thus clear that neither the bootstrap method nor the jackknife method can be applied indiscriminately. Care needs to be taken. For example, in the jackknife case the delete-one method does not yield a consistent estimate of the variance of the sample median. In the regression case, until this paper, the problem had remained unresolved. But for "inference purposes" both the nonparametric bootstrap method as well as jackknife method require the assumption that the error terms are independently and identically distributed with means 0 and constant variance. In addition, both require that at least $w_i = x_i^T(X^TX)^{-1}x_i \to 0$ as $n \to \infty$, a condition due to Srivastava (1971).

In Section 7, Wu provides a method for handling the unbalanced nature of the residuals for bootstrapping. However, it appears somewhat manipulative. Obviously, if the model is

$$Y_i^* = x_i^T\hat{\beta} + \frac{r_i}{(1-w_i)^{1/2}}t_i^*, \qquad i = 1\dots, n$$

(equation (7.1) of Wu), then the bootstrap estimator should be defined by

$$\hat{\beta}^*_{\text{new}} = (X^TD^{-1}X)^{-1}X^TD^{-1}Y^*, \qquad Y^* = (Y_1^*, \dots Y_n^*)^T,$$

where

$$D = \mathrm{diag}\left[(1-w_1)^{-1}r_1^2, \dots, (1-w_n)^{-1}r_n^2\right].$$

In the case $\theta = \beta$,

$$v_{*,\text{new}} = (X^T D^{-1} X)^{-1}.$$

The $v_{*,\text{new}}$ estimate of the variance should have less bias than most estimators suggested in the paper.

In the jackknife situation, Wu provides two sensible estimates of the variance, namely $\hat{v}_{J,r}$ and $\tilde{v}_{J,r}$, both employing an arbitrary scaling method, one externally and the other internally. However, $\tilde{v}_{J,r}$ does not perform well in his simulation studies. Probably the cause could be located had he considered pseudovalues, an idea he discarded in Section 5(a). Consider the pseudovalues $P_i = \hat{\theta} + C_1(\hat{\theta} - \hat{\theta}_{(i)})$, $i = 1, \dots, C$, where

$$C = \binom{n}{r}, \qquad r > k, \qquad C_1 = \binom{n-k}{r-k},$$

and $\hat{\theta}_{(i)}$ is the estimate $\hat{\theta}$ based on r observations. Recognizing the unbalanced nature, define the new jackknife estimate of θ by

$$\tilde{P} = \Sigma w_{(i)} P_i, \quad \Sigma w_{(i)} = 1$$

and

$$v_{J,r,} = \frac{1}{C_1 C_2} \Sigma w_{(i)} (P_i - \tilde{P})(P_i - \tilde{P})^T$$

$$= \frac{C_1}{C_2} \Sigma w_{(i)} (\hat{\theta}_{(i)} - \hat{\theta})(\hat{\theta}_{(i)} - \hat{\theta})^T,$$

where

$$C_2 = \binom{n-k}{r-k+1},$$

and the $w_{(i)}$'s are as in the paper. That is,

$$w_{(i)} = \frac{|X_{(i)}^T X_{(i)}|}{\Sigma_{i=1}^{c} |X_{(i)}^T X_{(i)}|}, \qquad i = 1, \dots, C,$$

where $X^T = (x_1, \dots, x_n)$ and $X_{(i)}$ is obtained by choosing any r vectors from $x_1, \dots, x_n$. Thus $X_{(i)}$ is an $r \times k$ matrix. For example, when $r = n - 1$, that is, the delete-one case, then for $\theta = \beta$,

$$P_i = \hat{\beta} + (n-k)(\hat{\beta} - \hat{\beta}_{(i)}),$$

where

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \qquad \hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}$$

and $X_{(i)}$ and $Y_{(i)}$ are obtained after deleting the ith observation. Letting

$$w_i = 1 - x_i^T(X^TX)^{-1}x_i,$$

we get

$$\begin{aligned}\left|X_{(i)}^T X_{(i)}\right| &= |X^TX - x_i x_i^T| \\ &= |X^TX|(1 - w_i)\end{aligned}$$

and

$$\sum_{i=1}^{n} |X_{(i)}X_{(i)}| = |X^TX|(n - k).$$

Hence,

$$\begin{aligned}\tilde{P} &= \hat{\beta} + \Sigma(1 - w_i)\big(\hat{\beta} - \hat{\beta}_{(i)}\big) \\ &= \hat{\beta} + (X^TX)^{-1}\Sigma x_i r_i \\ &= \hat{\beta},\end{aligned}$$

since $\hat{\beta} - \hat{\beta}_{(i)} = (X^TX)^{-1}x_i r_i/(1 - w_i)$, where r_i is the ith residual defined in the paper and $\Sigma x_i r_i = 0$, $V_{J(1)} = (X^TX)^{-1}\Sigma(1 - w_i)^{-1}x_i x_i^T r_i^2 (X^TX)^{-1}$.

Thus, this method yields the same $v_{J,r}$ as in Wu and additionally it provides an estimate of the bias by $C_1\Sigma w_i(\hat{\theta} - \hat{\theta}_{(i)})$. The estimator $\tilde{v}_{J,r}$ does not fit in this scheme.

In Sections 2 and 5, Wu mentions the strongness of the conditions used in Miller (1974). However, it should be mentioned that all the results with the exception of the asymptotic normality of $f(\hat{\beta})$, require only the condition that $w_i = x_i^T(X^TX)^{-1}x_i \to 0$, a condition first given by Srivastava (1968, 1971, 1972) in a series of papers. In these papers it is also shown that if $n^{-1}(X^TX) \to \Sigma > 0$ and $n^{-1/2}X \to 0$, then $w_i \to 0$. The asymptotic normality of $(X^TX)^{-1/2}(\hat{\beta} - \beta)$ requires only that $w_i \to 0$.

Turning to the simulation results reported in Tables 1–4, it would have been nice to know for which values of β_0, β_1 and β_2, the results in Table 1 are reported. While the variance estimates v_b, $v_{J(1)}$, $v_{J,8}$ and $v_{H(1)}$ involve only the residuals, the estimates v_J, v^* and v_w^* require that estimates be made of these parameters. The bootstrapping should have been done with more bootstrap samples rather than just 480. Efron (1986) recommends 1000, based on the coefficient of variation of the percentile. In Efron's notation

$$\mathrm{CV}\big[\hat{G}_B^{-1}(\alpha)\big] = \frac{\alpha(1 - \alpha)^{1/2}}{B^{1/2}\hat{G}^{-1}(\alpha)\hat{g}(\hat{G}^{-1}(\alpha))} + O(B^{-1}),$$

where $\hat{G}_B^{-1}(\alpha)$ is the Monte Carlo approximation to $G^{-1}(\alpha)$ based on B bootstrap samples. For $\hat{G}$ normal, CV[$\hat{G}_{1000}$ (0.95)] = 0.040. However, the bootstrap estimate $\hat{\theta}_b$ will not have a normal empirical distribution function and so a larger sample would have been desirable. Comparable to 1000 bootstrap samples is the number of jackknife subsets of 6, namely 924. Indeed as Wu himself points out, it would be worthwhile to compare results for a broad range of sizes of jackknife subsets.

From the simulation results presented in the paper, it seems the delete-one jackknife is the winner in overall performance, although considerably more simulation needs to be done before pronouncing it a clear winner. Also, in the unequal variance case a pure Monte Carlo simulation should have been performed in order to establish a standard of comparison for the results of Tables 3 and 4. These tables show that VLIN perform as well as Fieller's and better than the rest of the methods. However, in the unequal variance case, Fieller's method is not applicable. Although, it has been shown by Chan and Srivastava (1985) that Fieller's method is robust against certain departure from normality, and gives better results than the bootstrap method, it is not known how good the interval is in the unequal variance case. In the same paper it is also shown (equal variance case) that the VLIN method gives results comparable to Fieller's in the normal as well as in some nonnormal situations. A similar result was obtained in the case of sample variances in Srivastava and Chan (1985).

Finally, it would not be inappropriate to mention that most efforts have been devoted toward obtaining robust estimators of the variance of an estimate that should not be used under the circumstances described in the paper. Many regression diagnostic techniques should be able to detect any heteroscedasticity in the data. The estimates should accordingly be adjusted and the estimation of the variance of this adjusted estimate should have occupied more space. Nevertheless, I enjoyed reading the paper. The paper is thought-provoking and should lead to great activity in this area.

Acknowledgment. I wish to thank K. J. Keen for some discussion and a few suggestions on the computational aspect of the paper.

REFERENCES

CHAN, Y. M. and SRIVASTAVA, M. S. (1985). Robustness of Fieller's theorem and comparison with bootstrap method. Technical report 7, Univ. of Toronto.

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

HINKLEY, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics* **19** 285–292.

MILLER, R. G. (1974). An unbalanced jackknife. *Ann. Statist.* **2** 880–891.

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195.

SRIVASTAVA, M. S. (1968). On a class of nonparametric tests for regression parameters (abstract). *Ann. Math. Statist.* **39** 697.

SRIVASTAVA, M. S. (1971). On fixed-width confidence bounds for regression parameters. *Ann. Math. Statist.* **42** 1403–1411.

SRIVASTAVA, M. S. (1972). Asymptotically most powerful rank tests for regression parameters in MANOVA. *Ann. Inst. Statist. Math.* **24** 285–297.

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.