A little reflection, bearing in mind the conditioning argument, makes these statements seem almost tautological. In each case conditioning is on the sample, and the approximations are good up to terms of *smaller* order than $n^{-1/2}$. It is not true that the conditional distribution of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}$ is a good approximation to the distribution of $(\hat{\theta} - \theta)/\hat{\sigma}$. In this case the Edgeworth expansions of coverage probability for one-sided confidence intervals differ by terms of order $n^{-1/2}$. The same conclusion may be reached intuitively, noting that the statistic $(\hat{\theta} - \theta)/\sigma$ is not pivotal if σ is unknown. Work in Singh (1981), for example, concerns the approximation in (1) although I know of some authors who have tried to use it to promote an approximation of the distribution of $(\hat{\theta} - \theta)/\hat{\sigma}$ by that of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}$.

I should make one final remark to tie these comments to those made by Professor Wu prior to his formula (2.10). Since the conditioning in (1) and (2) is on the sample, then $\hat{\theta}$ and $\hat{\sigma}$ are effectively constant, and so the conditional distribution of $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}$ is just a location and scale change of that of $\hat{\theta}^*$.

REFERENCE

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

DEPARTMENT OF STATISTICS
AUSTRALIAN NATIONAL UNIVERSITY
GPO BOX 4
CANBERRA ACT 2601
AUSTRALIA

DAVID HINKLEY[1]

University of Texas at Austin

Professor Wu is to be congratulated for making a significant advance in jackknife methodology. The general use of information measures to determine weights in subsampling schemes is surely correct, and the implementation here for regression is most interesting.

The one somewhat negative conclusion of the paper concerns the comparatively poor performance of the bootstrap. It is to this that I shall address my remarks, because the bootstrap approach has, quite innocently, been misapplied. Good results *can* be obtained with bootstrap methods, as I hope to explain with the help of relatively simple examples.

The first point has to do with conditional probability, which in the regression context arises from conditioning on the experimental vector $\mathbf{x}$ of explanatory variables. The key issue can be seen most easily in the simple linear regression

[1]Research supported by National Science Foundation grant DMS-8505769.

model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Whatever the distribution of errors ε_i, it is standard practice to analyze the data conditional on $\mathbf{x} = (x_1, \ldots, x_n)$. For example, suppose that errors are assumed homogeneous and independent with variance σ^2. Then the variance of the OLS slope estimate b_1 is taken to be

$$\text{Var}(b_1|\mathbf{x}) = \sigma^2 / \sum (x_i - \bar{x})^2. \tag{1}$$

If the variances of the ε_i are σ_i^2, $i = 1, \ldots, n$, then we take the variance of the OLS slope to be

$$\text{Var}(b_1|\mathbf{x}) = \sum \sigma_i^2 (x_i - \bar{x})^2 / \left\{ \sum (x_i - \bar{x})^2 \right\}^2. \tag{2}$$

These two formulae are used even when the x's are obtained by random selection (independent of $\beta_0, \beta_1 \sigma^2$), as is justified by the conditionality principle (Kalbfleisch (1975)).

Formula (1) is directly estimated by the "homogeneous-errors" bootstrap, where random residuals are added to fitted values to obtain simulated responses at the experimental x-values. But the more general bootstrap, where *pairs* (x, y) are randomly sampled, does *not* estimate conditional quantities such as (1) and (2). Let me be more specific. Suppose that $\hat{G}$ and G represent, respectively, the joint empirical and true distributions of (x, y), and that bootstrap and real data, respectively, consist of n randomly sampled pairs (x_i^*, y_i^*) from $\hat{G}$ and G. Then the bootstrap estimate of variance for OLS slope b_1 is

$$\text{Var}(b_1; \hat{G}) \approx \text{Var}(b_1; G),$$

in clear contrast to what we want, $\text{Var}(b_1; G|\mathbf{x}^* = \mathbf{x})$. The contrast disappears in the limit as $n \to \infty$ (Freedman (1981)), but does not disappear in the data. An example will be given in a moment.

But can we hold $\mathbf{x}^* = \mathbf{x}$? Clearly not, if we sample from $\hat{G}$, because we would then always get the same responses y (unless there is replication in the data). But it is not necessary to hold $\mathbf{x}^* = \mathbf{x}$ in order to estimate $\text{Var}(b_1; G|\mathbf{x}^* = \mathbf{x})$. For example, (1) depends on $\mathbf{x}$ only through $A = \Sigma(x_i - \bar{x})^2$. A simple approach would be to partition the bootstrap samples according to interval values of $A^* = \Sigma(x_i^* - \bar{x}^*)^2$, and to then estimate $\text{Var}(b_1; G|\mathbf{x}^* = \mathbf{x})$ from those samples with A^* in the interval that includes A itself. Of course A is the most relevant partition statistic only if the errors are homogeneous, and it is not obvious what to do in general. But before we get to that complication, let me show that the simple partition method works.

A single sample of $n = 19$ responses $y = x + \varepsilon$ was generated, the ε's being $N(0, 1)$ numbers generated by IMSL subroutine GGNML, and the x-values being 1(1)5, 13, 20(1)25, 33, 40(1)45. The OLS fitted equation was $0.17 + 0.994x$, and the usual standard error calculation for $b_1 = 0.994$ gave the result 0.0144. Then $N = 1000$ bootstrap samples of size 19 were generated by random sampling of (x, y)'s. The bootstrap standard error for b_1 was 0.0167. Values of A^* ranged

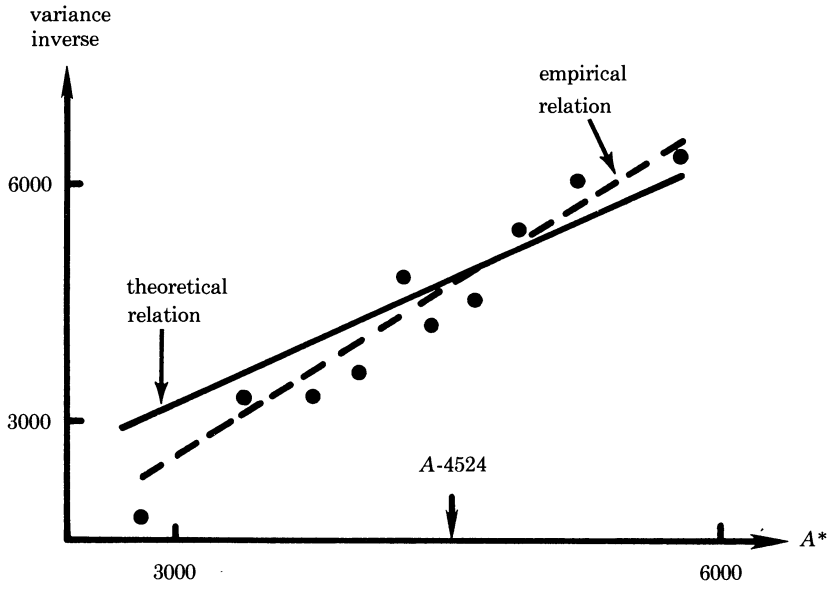

FIG. 1. *Conditional bootstrap variance reciprocals, for OLS slope estimate in simple linear regression of y on x, graphed against ancillary statistic* $A^* = \Sigma(x^* - \bar{x}^*)^2$. *Each point obtained from* 100 *bootstrap samples.*

from 2500 to 6000, and the data value of A was 4524. The bootstrap samples were partitioned into ten groups each of 100 samples on the basis of intervals for A^*, and standard errors for b_1 were calculated separately for each group. The results are plotted as variance inverses in Figure 1, the abscissa being the average A^* value for each interval; the solid line on the figure represents the "theoretical" result

$$\text{Var}(b_1; G|\mathbf{x}^* = \mathbf{x}) = \hat{\sigma}^2/A^*.$$

Smooth interpolation of the standard error for b_1 at $A^* = A$ gives the result 0.014—almost exactly the right answer.

Now what of the general case? One appropriate strategy is to use an empirical analog of (2) for A, rather than (1). The result may be unstable if the σ_i^2 are replaced by individual squared residuals. Another strategy is to use the same A as before, and this will often work well because of its high correlation with the "best" partition statistic.

There is not space here to discuss this in adequate detail, but when we are estimating confidence probabilities for β_1 it is possible to use quite sophisticated techniques to achieve the appropriate conditioning without an explicit partition of the bootstrap samples. Further discussion and detailed examples can be found in Hinkley (1986) and Hinkley and Schechtman (1985).

It seems very likely that the lack of conditioning has an appreciable effect on Professor Wu's example in Section 10. Perhaps more serious, however, is the

failure to work with a pivot in defining the bootstrap confidence interval procedure.

To understand the relevance of pivots, consider the simple problem of setting confidence limits for a mean μ using a sample average $\bar{x}$. One approach is to begin by estimating percentiles of $\bar{x} - \mu$, which is to say estimating the quantities $d(p, F)$ that satisfy

$$\Pr(\bar{x} - \mu \leq d(p, F)) = p,$$

where F is the true distribution of x's. When the standard bootstrap is used, we estimate $d(p, F)$ by $d(p, \hat{F})$. Then an equitailed $1 - \alpha$ confidence interval for μ is $(\bar{x} - d(1 - \frac{1}{2}\alpha, \hat{F}), \bar{x} - d(\frac{1}{2}\alpha, \hat{F}))$. This usually does not work, because $d(p, F) \neq d(p, \hat{F})$. But if we replace $\bar{x} - \mu$ by a pivot, then the corresponding percentiles are, in principle, known—because by definition the distribution of a pivot does not depend on unknowns. In the parametric case, possible pivots are (i) $\log_e \bar{x} - \log_e \mu$ for scale families such as gamma distributions and (ii) $\sqrt{n}(\bar{x} - \mu)/s$ with s^2 the sample variance for location-scale families such as normal distributions.

The general concept and identification of (approximate) pivots is a little more complicated in the nonparametric case (Chapman (1985); Chapman and Hinkley (1986)). In certain rather restricted cases, Efron's percentile method bypasses the need for pivots. One simple but useful general approach is to always studentize

$$\text{estimate} - \text{parameter},$$

for example using some type of jackknife standard error for the estimate. That is, if T estimates θ, and if SE is the estimated standard error of T, then we would bootstrap $Z = (T - \theta)/SE$ to obtain percentile estimates $\hat{d}_Z(p)$ for Z, leading to confidence limits $(T - SE \cdot \hat{d}_Z(1 - \frac{1}{2}\alpha), T - SE \cdot \hat{d}_Z(\frac{1}{2}\alpha))$.

By way of illustration, again for the case of the mean, let me quote some numerical results from Chapman (1985). Data samples of size $n = 30$ were generated from the χ_3^2 distribution. For each data sample, the bootstrap was applied to $\bar{x} - \mu$, $\log_e \bar{x} - \log_e \mu$ and $(\bar{x} - \mu)/s$, and corresponding 90% confidence intervals were then calculated for μ. Efron's percentile method was also applied. In principle, the true μ should fall outside the interval in 10% of the cases—5% of the time on the left and 5% of the time on the right of the interval. Table 1 shows the actual left, right and total error rates for 3000 data sets.

It is quite easy to see that the quantity $\hat{\theta} - \theta$ in Professor Wu's example of Section 10 is far from being a pivot. For example, the normal-theory variation of $\hat{\theta}$ is quadratic in θ. It would be interesting to see the improvement in bootstrap performance if $\hat{\theta} - \theta$ were standardized using a nonparametric delta-method estimate of standard error (Efron (1981)), and if conditional confidence for the actual $\mathbf{x}$ were estimated.

In conclusion, I hope that my comments will indicate that bootstrap methods can be rendered more effective by the use of those general principles that make standard methods work when the latter are appropriate. This is not necessarily easy advice to follow, and a good deal of further theoretical research is needed.

TABLE 1

Estimated error rates for bootstrap confidence limits on mean, from analysis of 3000 data sets of size $n = 30$; bootstrap applied with $B = 499$. [Source: Chapman (1985).]

Method	Error rates %		
	Left	Right	Total
Bootstrap $\bar{x} - \mu$	11	3	14
$(\bar{x} - \mu)/s$	6	5	11
$\bar{x}/\mu$	6	6	12
Efron's percentile method	9	5	14
Exact	5	5	10

REFERENCES

CHAPMAN, P. (1985). Ph.D. thesis, Univ. of Minnesota.

CHAPMAN, P. and HINKLEY, D. V. (1986). The double bootstrap, pivots and confidence limits. Technical Report 28, Center for Statistical Sciences, Univ. of Texas.

EFRON, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* **68** 589–599.

FREEDMAN, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9** 1218–1228.

HINKLEY, D. V. (1986). Constructed variables and transformation diagnostics. Technical Report 26, Center for Statistical Sciences, Univ. of Texas.

HINKLEY, D. V. and SCHECHTMAN, E. (1985). Conditional bootstrap methods in the mean-shift model. Technical Report 25, Center for Statistical Sciences, Univ. of Texas.

KALBFLEISCH, J. D. (1975). Sufficiency and conditionality (with discussion). *Biometrika* **62** 251–268.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF TEXAS
AUSTIN, TEXAS 78712

THOMAS MITCHELL-OLDS[1,2]

University of Wisconsin-Madison

The resampling procedures discussed by Professor Wu provide an important solution to several problems of current interest to population geneticists. Measuring natural selection in wild populations of plants and animals has long been

[1]Present address: Department of Botany KB-15, University of Washington, Seattle, Washington 98195.

[2]Supported by National Science Foundation grant BSR-8421272 and a National Institutes of Health Postdoctoral Traineeship in genetics.