TABLE 2
*Percent coverage for nominal 90% confidence interval**

	Bootstrap method							
	$\sigma = 0.1$				$\sigma = 0.5$			
X	T1	T1 BC	T2	T2 BC	T1	T1 BC	T2	T2 BC
0	74	94	50	50	83	88	65	63
1/8	0	7	81	50	39	76	70	67
1/4	100	100	100	94	96	89	90	90
3/8	0	3	74	48	33	76	67	66
1/2	100	100	100	95	97	90	91	90

*T1 means type I percentile bootstrap; T1 BC is type I percentile bias-corrected; T2 is type II bootstrap; T2 BC is type II bias-corrected as described in text.

To confuse matters further, in using the bootstrap to pick a bandwidth for a kernel density estimate, the model generating bootstrap data must be an oversmoother. Failure to recognize these subtleties will result in very poor inferences.

Applying the smoother and then the bootstrap is a breeze (our simulations caused the breeze to blow 1,000 times), and we were able to commit malpractice with practically no effort. There are many other examples where hidden problems with the bootstrap will occur unless one is especially knowledgeable and careful.

Our response to Young's paper and to our example is a call to action. The statistical profession needs to communicate the good news, the bad news and the "no news yet." The bootstrap will succeed for a broad class of models and data structures. It will fail in others; sometimes it can be rescued by modifications that attend to the structure of the problem. We need to communicate what we know about the procedure's strengths and weaknesses and to identify situations where we do not yet know the answers. This communication must reach current and potential users and thus must appear in a broad array of journals and other information sources. As we learn more, information needs to be updated. Of course, the same recommendations hold for all statistical procedures, but the attraction of the bootstrap makes the need most acute.

ACKNOWLEDGMENTS

Partial support was provided by Grant RO1 CA51932 from the National Cancer Institute.

Comment

David Hinkley

INTRODUCTION

This is a timely article. It is likely to appear in print about the same time as first reviews of the excellent introductory book by Efron and Tibshirani (Efron and Tibshirani, 1993), a book which should allay some of the impatience and scepticism that I sense in the sophisticated user community about the bootstrap as a practical tool. We are also beginning to see the first wave of software products which claim to do bootstrap analysis: some of these are embarrassingly naive. Let us hope for more good applications-oriented books and better software products.

I think that Alastair Young has done an excellent job of highlighting the key theoretical developments and has suggested some sensible steps for further research. Much of what I have to say will complement his assessment and will focus on a few practical points.

WHEN DOES BOOTSTRAP WORK?

This question comes up twice in the paper, in the context of nonparametric bootstrapping of a point estimator. The first time we are given a succinct mathematical characterization which is clearly useless to even the best applied statistician. The second time

David Hinkley is Professor of Statistical Science and Head of the Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, England.

Young suggests that there may be a purely empirical answer, and this is of course what we need—and what we may not get without a shift in research emphasis.

What does the question actually mean? Some of the issues are as follows. Would the particular bootstrap method work on a very large, clean data set with infinite simulation? Is there degradation because of finite simulation? Is the data set too small for good nonparametric analysis? Is contamination in the data causing problems? Are there outliers in the bootstrap output because of outlying bootstrap samples? Should another statistic be used? Should a different bootstrap method be used? And so on.

It seems clear that both experience and detailed assessment of data should be brought to bear. For example, it would seem silly to bootstrap a correlation rather than its z-transform. For regression problems the usual diagnostics should be run, in part to try to characterize the nature of random variation: restricted randomization should be considered, so that certain patterns of simulated data sets are avoided. (Where is Fisher when we need him?)

One should be a bit cautious about trusting asymptotic theory, because it deals with potential rather than reality. For example, early theoretical work suggested that the simple bootstrap would work for the sample median: it gives a consistent estimate of variance, which the jackknife does not. In fact the bootstrap is poor at approximating the variance of the median in small ($n = 20$) clean data sets, and results for bootstrap confidence limits are bad. As later theory has shown, addition of small random perturbations to bootstrap samples is necessary to get good results for the median, but even now we do not know for sure exactly what prescription to offer.

To answer the question "When does bootstrap work?" we need first to answer the question "What does bootstrap do?" As far as bootstrapping a parameter estimate goes, it can be useful to examine how bootstrap results compare to results from more traditional analyses such as likelihood analysis. Here I should like to look at two small examples in order to make a rather simple but important point.

EXAMPLE 1. The following ordered sample of $n = 12$ lifetimes, taken from Cox and Snell (1981), is the ninth of 10 samples which are taken to be from different gamma distributions; we are interested in the mean:

$$3, 5, 7, 18, 43, 85, 91, 98, 100, 130, 230, 487.$$

What we shall do is compare parametric likelihood analysis with nonparametric bootstrap analysis. First the ML estimate of the mean μ is the sample average $\bar{y} = 108.0833$, and the ML estimate of the gamma index is $\widehat{\kappa} = 0.71$. However, a likelihood confidence interval for the index would easily include the value $\kappa = 1$. In isolation these data would be thought quite consistent with an exponential distribution, and on grounds of parsimony this model might be used in further analysis.

Now in a standard profile likelihood analysis for μ under the full gamma model we use the loglikelihood ratio $\mathrm{LLR} = L(\widehat{\mu}, \widehat{\kappa}) - L(\mu, \widehat{\kappa}_\mu)$, where

$$L = n\kappa(\log\kappa - \log\mu) + (\kappa - 1)s_0 - \kappa s_1/\mu - n\log\Gamma(\kappa)$$

is the loglikelihood with $s_0 = \Sigma \log(y_j)$, $s_1 = \Sigma\, y_j$ and $\widehat{\kappa}_\mu$ denotes ML of κ with μ fixed. An approximate 95% interval for μ is the set of values of μ for which $2\mathrm{LLR} \leq 3.84$. For our data, this interval for μ is $(57.1, 242.8)$; this is wider than under the best-fitting gamma model, $(59.1, 230.1)$, and very different from the interval $(64.45, 201.76)$ obtained under the exponential model. (I have not used exact methods for the last two intervals.)

For the nonparametric bootstrap, 999 samples were generated by random resampling. Figure 1 shows Q–Q plots of the ordered bootstrap sample means versus quantiles under the exponential and gamma models; the dotted lines correspond to theoretical expectation. The results are clear: nonparametric agrees with fitted gamma, and both are quite different from exponential. This is interesting, if predictable. But what about the confidence interval? To obtain an interval from the bootstrap results we apply the next-to-simplest method on the log scale, so the limits are calculated as

$$\exp(2\log(\widehat{\mu}) - \widehat{\mu}^*_{(r)}), \quad r = 25, 975,$$

where $\widehat{\mu}^*_{(r)}$, $r = 1, \ldots, 999$, are the ordered bootstrap mean estimates. (This is probably the best nonparametric bootstrap method for this problem.) The confidence interval is $(61.4, 246.4)$, more like the profile gamma likelihood interval than the exponential likelihood interval. Virtually the same interval would be obtained with the same pivot and its distribution under the fitted gamma model.

The lesson drawn from this is that bootstrap results will tend to mimic parametric model results under the best-fitting parametric model, *not* the simplest model which fits the data.

EXAMPLE 2. A more complex application to consider is regression. This is interesting because there are several bootstrap algorithms, ranging from unconditional resampling of cases to unconditional resampling of errors, with several options in between,

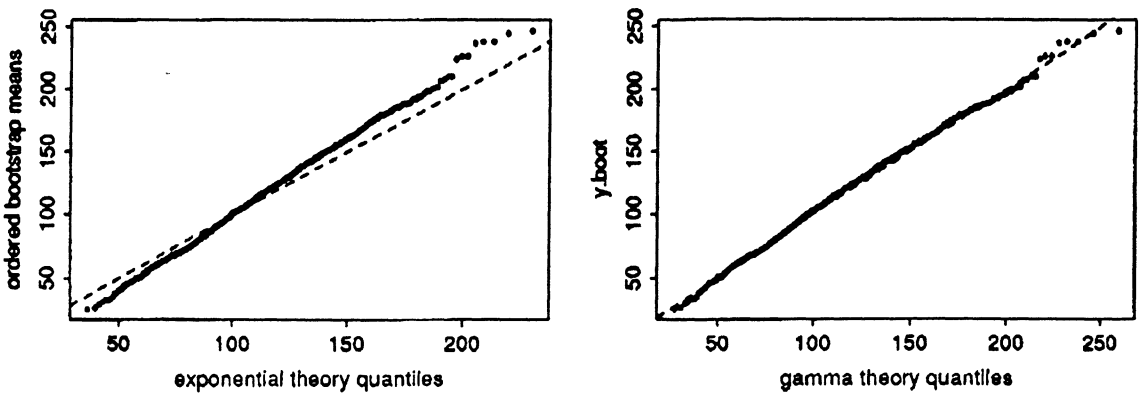

FIG. 1. *Q–Q plots of* 999 *bootstrap sample means: left and right panels correspond to theoretical quantiles under exponential and gamma model fits.*

including stratified resampling of cases and local resampling of errors (wild bootstrap, etc.). It would seem that a simple protocol could be laid down which is tied to an initial inspection of the design, the nature of the fitted model and residual-based diagnostics. This is what the applied statistician needs. How do the basic algorithms behave? There is an extended theoretical treatment of this by Wu (1986), but I know of no comparably comprehensive empirical study.

A simple example is provided by the data plotted in the left panel of Figure 2, a double-logarithmic plot of the body (x) and brain (y) weights of $n = 62$ mammals.

A fairly standard diagnostic analysis suggests that the simplest model, straight-line regression with homoscedastic random errors, is adequate. The point corresponding to humans (labelled) might be viewed as an outlier and so will be omitted from further analysis just to be safe. Then the slope estimate is $b = 0.742$ with estimated standard error $\mathrm{SE}(b) = 0.027$ according to the usual formula. The bootstrap which simulates data by adding random residuals (after suitably standardizing and centring) to the fitted model produces a nice normal plot of bootstrap slopes, agreeing with the model-based mean and standard error, as expected. However, the bootstrap which simulates data by randomly resampling cases does not give similar results. The right panel of Figure 2 shows the normal Q–Q plots of the slope estimates from 999 bootstrap samples. The solid line on the plot corresponds to the normal distribution with the usual standard error, that is, 0.027, while the dotted line corresponds to the robust standard error 0.020 obtained from the robust variance matrix $(X^TX)^{-1}X^TSX(X^TX)^{-1}$, in which S is the diagonal matrix of squared residuals (adjusted for leverage).

So the bootstrap sampling algorithm which does not assume an error model gives results which are robust against heteroscedasticity. We should really expect this: it is another version of the phenomenon of Example 1. But is this the kind of behavior we want, or would we like the bootstrap to produce results corresponding to the most parsimonious model consistent with the data? If the latter, then these two examples show that we must explicitly inject that model.

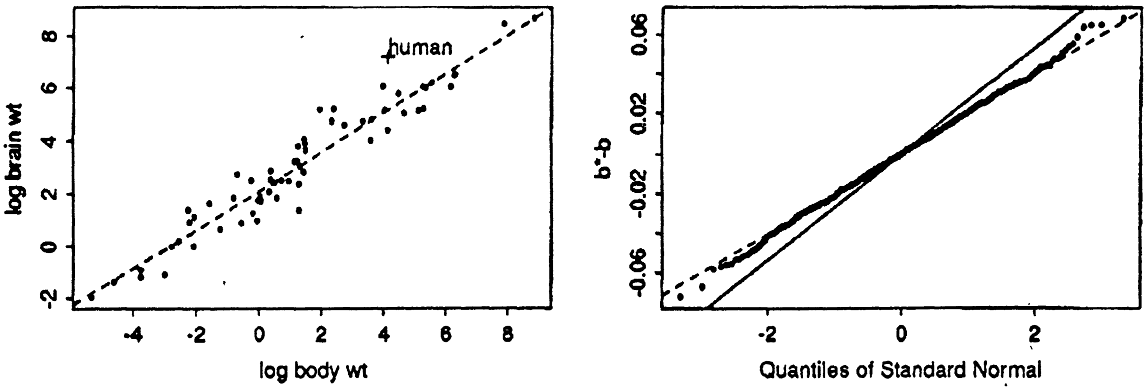

FIG. 2. *Regression example: left panel is log–log plot of data; right panel is normal Q–Q plot of* 999 *bootstrap slope estimates under random sampling of cases; solid and dotted lines correspond to usual and robust standard errors.*

DEALING WITH BOOTSTRAP OUTPUT

Accounts of bootstrap methodology are remarkably short on practical details. By this I mean advice on such things as the following:

- how to monitor output so as to know when to stop in a big problem where simulation is expensive;
- what to do with bootstrap outliers (is it OK to eliminate very deviant simulated estimates?);
- whether or not it is worth smoothing the bootstrap simulation output;
- what bootstrap diagnostic plots should be made and how they should be interpreted.

The area of diagnostic plots seems to me to be of considerable potential importance. The "jackknife-after-bootstrap" methodology is a very good start. In regression it can also be helpful to plot coefficient estimates against design summaries when whole cases are resampled.

ITERATED BOOTSTRAP

The reliability of a bootstrap algorithm can be assessed by bootstrapping the bootstrap. This sounds a bit exotic, but it can work in a variety of ways. For example, there is some very impressive theory by Beran, Hall and others to show that it can lead to useful corrections to confidence limits and other forms of bootstrap output. It is also possible to see how the distribution of a statistic might change with changes in parameter value, but I suspect that we need to take a much wider, method-oriented view of what iterated bootstrap has to offer. For example, there might be many useful diagnostic plots, some addressing the crucial questions "Does bootstrap work?" and "How should I modify the bootstrap to make it work?" Some restraint is needed because computation can become excessive, but there are often shortcuts (computational or methodological) which enable fast execution of iterated bootstrap. One shortcut that works for some problems is the "jackknife-after-bootstrap" algorithm.

APPLICATIONS

Young is right to refer to time series and stochastic process models as important, because for such models the bootstrap might be able to provide the answers that other methodologies cannot. There is certainly growing interest among econometricians in applying bootstrap ideas to such things as cointegration models; but, given the results of the excellent work done so far on time series problems, we must not expect miracles in the shape of completely general methods.

Another important area where nothing else is likely to work well is error rate estimation for general classification methods (trees, neural networks etc.) where accurate error rates are needed.

Model selection problems can surely benefit from bootstrap, although not in its simple form. Some useful work has been done already on subset bootstrap, a topic which probably deserves more attention.

One large topic to which we know the bootstrap can contribute much is nonparametric curve and surface fitting. There is already some excellent theoretical work on the topic. What I would hope for here is a portfolio of model-assessment tools which address questions about curve shape, interaction of effects and so forth. Such questions are almost impossible to handle without the bootstrap.

One general point to bear in mind is that one should know what one is bootstrapping. To me this means, where possible, expressing quantities in terms of distributions and estimated distributions, that is, making a precise theoretical definition of the problem. I think that Efron illustrates this beautifully in his treatment of missing data problems. I am not sure that the point has been fully taken on board in some of the work on stepwise regression fitting and on nonparametric curve fitting.

PARAMETRIC BOOTSTRAP

I think that I am not alone in sometimes wondering why there is a parametric bootstrap, except as a pedagogical device, but clearly there are strong advocates for use of, say, the ABC method of calculating confidence intervals for parametric models. These methods do undoubtedly serve a very useful theoretical purpose, not least in giving us the empirical exponential family; perhaps their development has acted as a spur to the developers of small-sample likelihood methods, so that use of Barndorff–Nielsen's famous r^* might soon become widely practicable. Then we shall need to know which of r^* and ABC is better for practical purposes, if indeed they differ much at all.

CONCLUSION

Bootstrap was a wonderful invention and has led to much fascinating and surprising research. As with all statistical methodology, a lot of difficult theoretical work has been, and continues to be, necessary to help validate the methodology. But theory alone merely gives pointers, and these are only useful if the right questions were asked. The methodology must be seen to provide empirically reliable new solutions and must drive the new theoretical questions.