1979a; Freedman, 1981):

1. Compute $\widehat{\beta} = (X^T X)^{-1} X^T Y$.
2. Let $e_1, \ldots, e_n$ be the residuals $e - Y - X\widehat{\beta}$.
3. Let $e_1^*, \ldots, e_n^*$ be an i.i.d. sample from the empirical distribution of $e_1, \ldots, e_n$.
4. The bootstrap model is $Y^* = X\widehat{\beta} + e^*$.

The bootstrap model is much like the real model, with the advantage that the "true" value of β, namely, $\widehat{\beta}$, is known. The bootstrap model works for inference about the distribution of $\widehat{\beta}$ in that if $\widehat{\beta}^* = (X^T X)^{-1} X^T Y^*$, then, under mild conditions on the rate of growth of the elements of X, the asymptotic distribution of $(\widehat{\beta}^* - \widehat{\beta})$ is the same as that of $(\widehat{\beta} - \beta)$ [see Freedman (1981)]. It might be hoped that this would enable the bootstrap model to reflect accurately the behavior of estimates based on selected columns of X as well. Unfortunately, this does not seem to be the case. Roughly speaking, this is because $\mathbb{E}\|X\widehat{\beta}\|^2 > \|X\beta\|^2$, while $\mathrm{Var}(e_i^*) = (1/n)\sum_{i=1}^n \mathbb{E}(e_i^2) = (n-p)\sigma^2/n$. In other words, the mean of Y^* tends to be larger than that of Y, while its variance tends to be less. Thus the bootstrap model tends to confirm the overoptimistic assessment of goodness of fit produced by model selection. The asymptotic performance of the bootstrap is good as $n \to \infty$ with p fixed, since $(1/n)\|X\widehat{\beta}\|^2 \to (1/n)\|X\beta\|^2$ under mild conditions on the rate of growth of the elements of X. When p is a substantial fraction of n however, which is often the case in variable selection, the results can be quite misleading (Freedman, Navidi and Peters, 1988). Potential solutions may involve shrinking the length of $\widehat{\beta}$ for use in the bootstrap model. Since the use of model selection procedures is quite extensive in statistical practice, better methods of assessing the performance of selected models would be very useful. I think it is likely that the bootstrap will turn out to have something to offer in this area.

Comment

Mark J. Schervish

Professor Young is to be congratulated on summarizing so succinctly and clearly the vast body of work on the bootstrap which has appeared since 1979. The bootstrap has achieved a remarkable level of notoriety both due to its analytical simplicity and to its seeming ability to serve up the proverbial "free lunch." However, behind all of the technical details of the bootstrap and its asymptotics, there still lies the question of why does (or does not) the bootstrap work in general. The theoretical use of the bootstrap involves the replacement of a distribution F in a formula $T(X, F)$ by some other distribution $\widehat{F}$. The degree to which this replacement is successful depends on the degree to which $\widehat{F}$ resembles F in important regards. For example, suppose that F is a distribution with finite variance, $\widehat{F}$ is the empirical distribution and $T(X, F)$ is the average $\overline{X}$ of the sample X minus the mean of the distribution F. Then the variance of $T(Y, \widehat{F})$ (where Y is a sample from $\widehat{F}$) can be expected to be a lot like the variance of $T(X, F)$. On the other hand, if F is a continuous distribution on an interval $[0, \theta]$ and $T(X, F) = n(\theta - X_{(n)})$, where $X_{(n)}$ is the largest order statistic, then Young points out the well-known fact that $\Pr(T(Y, \widehat{F}) = 0)$ converges to $1 - \exp(-1)$ as $n \to \infty$, while $T(X, F)$ has a continuous distribution.

I believe that some insight into what the bootstrap does can be gained by doing something with this last example that is uncommon in most bootstrap applications, namely, that we think about the problem. An obvious observation is that $\widehat{F}$ and F differ markedly in the manner in which the largest order statistic from a sample is related to the least upper bound on the support of the distribution. In particular, with $\widehat{F}$, the two can be equal with nonnegligible probability; with F, they cannot. An obvious, albeit naive, response is to smooth $\widehat{F}$, that is, replace the empirical distribution by a continuous distribution which approximates it. For example, if $X_{(1)} \le \cdots \le X_{(n)}$ are the order statistics, one could define $\widehat{F}(x) = G(u)i/n + [1 - G(u)]\ (i-1)/n$ for $X_{(i-1)} < x < X_{(i)}$, where G is a continuous distribution function and $u = (x - X_{(i-1)})/(X_{(i)} - X_{(i-1)})$. (Forget about $x < X_{(1)}$ for now.) Bickel and Freedman (1981) claim that even this does not mend the problem. They attribute (page 1210) the problem to "the lack of uniformity in the covergence of" $\widehat{F}$ to F. In fact, it is not difficult to see what happens in this case. We get that $T(Y, \widehat{F})$ is the sum of two random vari-

Mark J. Schervish is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

ables, one of which is precisely the same as when $\widehat{F}$ was the empirical distribution, and the other has the distribution of $(1-U)Z$, where U and Z are independent and U has distribution G and Z has the distribution of $n/X_{(n)}$ times the difference between two successive order statistics corresponding to $Y_{(n)}$ and the $X_{(i)}$ which comes immediately before $Y_{(n)}$. For each fixed k, the distribution of $n(X_{(n-k+1)} - X_{(n-k)})/X_{(n)}$ is asymptotically the same as that of $T(X, F)$. In other words, at least the smallest values of $T(Y, \widehat{F})$ should look a bit more like they come from the correct distribution. Although the poor behavior is not completely remedied, a little thought brings us a little closer to the solution we sought, or so it seems.

Perhaps the problem which the bootstrap suffers lies deeper than what we have discovered so far. A fundamental flaw in the bootstrap logic is that one is asked to replace F by $\widehat{F}$ without taking into account the uncertainty which remains concerning F. In the example at hand, one is particularly hamstrung by not being able to account for the uncertainty about θ, let alone the uncertainty about he distribution of X given θ. For example, suppose that we pretend to be a Bayesian, but a sloppy Bayesian who is trying to mimic the bootstrap. This sloppy Bayesian might decide to use a Dirichlet process prior (see Ferguson, 1973) for the distribution of F. Unfortunately, the least upper bounds of all of the F's which arise from a Dirichlet process are the same. One could then model F as having a Dirichlet process distribution with base measure α_θ given θ, where α_θ is concentrated on $[0, \theta]$. Mechanically, the sloppy Bayesian will simulate a θ from the posterior distribution and then simulate a sample Y from the conditional posterior of F given θ and then calculate $n(\theta - Y_{(n)})/\theta$ and repeat. Of course Dirichlet processes are discrete with probability 1, but if the base measure does not put positive mass on θ, then with probability 1 the largest order statistic will be less than θ. The funtion $c(\theta) = \alpha_\theta([0, \theta])$ governs how the rate at which repeats appear depends on θ. If $c(\theta)$ is chosen to be constant and α_θ has a density a_θ, then the likelihood function for θ becomes $\prod_{i=1}^{n} a_\theta(X_i)$, for $\theta \geq X_{(n)}$. As Barron (1986) points out in his comment on the use of a similar prior distribution by Diaconis and Freedman (1986), this has the effect of treating the data as if it were a random sample from a distribution with density a_θ/c as far as inference about θ is concerned. Suppose, for example, that $a_\theta(x) = 1/\theta$ for $0 \leq x \leq \theta$, but the data comes from a standard normal distribution truncated to the interval $[0, \theta]$. The data distribution is less likely to produce observations close to θ than the uniform, hence the posterior distribution of θ will be too highly concentrated near $X_{(n)}$. Another problem with the Dirichlet process is that, because it is discrete, multiple repeated values will appear in Y much more often than in the bootstrap sample. Put another way, the probability that $Y_{(n)}$ is small compared to $X_{(n)}$ is larger for the Dirichlet than for the bootstrap. An ad hoc procedure might be to combine the Dirichlet and bootstrap analyses by using the posterior distribution for θ as the sloppy Bayesian does, but then sample the Y data as the bootstrap does. I have simulated a number of data sets and applied the four methods already described to each of them. Some results are summarized in Figure 1. The simulated data arose from standard normal random variables conditional on being in the interval [0, 1.5]. Two hundred X-samples of size n of such data were simulated and, for each sample, one hundred Y-samples were simulated. For each X-sample, the 100 values of $n(\theta - Y_{(n)})/\theta$ were then sorted from smallest to largest, and then these 200 sets of order statistics (one for each X-sample) were averaged to produce the values on the vertical axes. The horizontal axes are the $i/101$ quantiles of the Exp(1) distribution for $i = 1, \ldots, 100$. The straight line has slope 2.2298, because the asymptotic distribution of $(1.5 - X_{(n)})/1.5$ is Exp(1/2.2298). (The prior distribution for θ was a continuous mixture of uniform on [0, 1] and Pareto with density proportional to $1/\theta^2$ for $\theta > 1$.) Ideally, the points should lie on the straight line. We see the effect of $Y_{(n)}$ being too small on the Dirichlet analysis as well as the effect of $Y_{(n)} = X_{(n)}$ so often in the bootstrap analysis. We also see the improvement which smoothing provides in the lower tail of the distribution. What might be surprising is how well the ad hoc combination of bootstrap sampling with sampling from the posterior of θ seems to do. By thinking a little more about what was important in this problem, we appear to have made

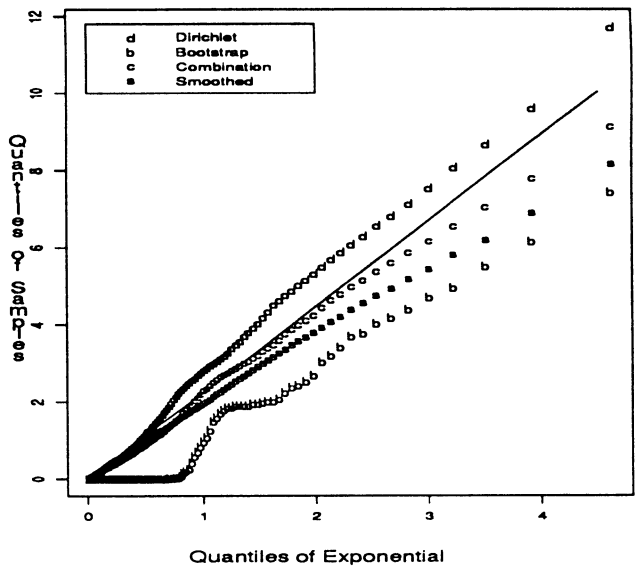

FIG. 1. *Sample distribution of* $(\theta - Y_{(n)})/\theta$ *with* $\theta = 1.5$ *and* $n = 50$.

some additional progress, but even this assessment may be premature.

Emboldened by this slight degree of success, the sloppy Bayesian might try to overcome some of the problems inherent in the Dirichlet process prior. There exist classes of tail-free processes (see Ferguson, 1974) which are absolutely continuous with respect to Lebesgue measure with probability 1. Mauldin, Sudderth and Williams (1992) and Lavine (1992) describe a subclass of these processes (called Polya trees) for which numerical calculations are feasible. Suppose that our sloppy Bayesian tries to be a little less sloppy and wishes to use one of these Polya tree priors for which densities exist with probability 1. Naively, one could proceed just as with the Dirichlet. First, sample θ from the posterior of θ and then sample Y from the conditional posterior of F given θ, calculate $n(\theta - Y_{(n)})/\theta$ and repeat. One would quickly discover that $n(\theta - Y_{(n)})/\theta$ tends to be very large, and the posterior distribution of θ is much more spread out than one would anticipate from knowledge of uniform distributions and other distributions with positive density on bounded intervals. This behavior lies at the heart of the problem being solved. Even though all of the observed data lies in an interval (such as [0, 1.5] in the example), there is significant probability that θ is very large. The reason is that the class of distributions with positive density on bounded intervals includes many distributions whose densities, although positive on the entire interval $[0, \theta]$, are incredibly small near θ. The very idea of calculating $T(X, F) = n(\theta - X_{(n)})/\theta$ is founded on the assumption that the density of F does not go to 0 at θ. Otherwise, $T(X, F)$ does not have a nondegenerate asymptotic distribution. For example, if the density of F approaches 0 linearly as $x \to \theta$, then $\sqrt{n}(\theta - X_{(n)})/\theta$ has a nondegenerate asymptotic distribution. One should not be surprised to see $n(\theta - Y_{(n)})/\theta$ being very large when the densities can drop to near zero at θ. In fact, this observation calls into question the very problem of trying to estimate the distribution of $n(\theta - X_{(n)})/\theta$ without further assumptions on F. For example, should we assume that the density of F does not go to 0? Should we assume that there is a uniform lower bound on all densities under consideration? We could go on and on about how to solve more focused problems with more carefully thought out assumptions, but that would take us too far from the topic of this discussion. The point is that the original problem is ill-posed, but we did not recognize this fact until we thought the problem through more carefully.

In summary, what do we learn from all of these failures to solve what appeared to be a fairly straightforward problem? First, we should learn that "automatic" approaches to inference, like the bootstrap, are dangerous because they discourage thinking about the problem. It is not the use of the computer to replace analysis which is the danger in using the bootstrap, but rather the misconception that serious thought about underlying assumptions can be replaced by pretending that $\widehat{F}$ is close enough to F. For the dependent data problems described by Professor Young in Section 5, more assumptions need to be made than in the independent data case, and more thought must be given to each problem. This seems to be movement in the correct direction and should be encouraged. Second, even when replacing F by $\widehat{F}$ in a theoretical calculation and then blasting ahead is fine for an asymptotic analysis, statisticians should never lose sight of the fact that, with finite data, uncertainty remains about everything we do not know. Even if we can convince ourselves that $\widehat{F}$ has many of the important properties of F in which we are interested, we still need to take care to say how uncertain we are about replacing F by $\widehat{F}$. Personally, I think that this is where serious research on the bootstrap ought to be undertaken. For example, in problems like confidence intervals, where there is some agreement that the bootstrap has been successful, how should one sensibly express the degree of uncertainty which remains concerning F? Some of the Bayesian bootstrap literature [e.g., Rubin (1981), Lo (1987) and Banks (1988), to name a few] makes a very small step in this direction, but much more is needed. What we do not need to learn from the example described above is that the reason the bootstrap fails is that the problem is ill-posed. The reason the bootstrap fails in this problem, and in others where it appears to succeed, is that it hinders us from discovering the nature of the problem by discouraging thought. Until this flaw is overcome, it will be difficult if not dangerous to take the bootstrap seriously as a statistical tool.