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.