Bootstrap methods are a collection of sample re-use techniques designed to estimate standard errors and confidence intervals. Making use of numerous samples drawn from the initial observations, these techniques require fewer assumptions and offer greater accuracy and insight than do standard methods in many problems. After presenting the underlying concepts, this introduction focuses on applications in regression analysis. These applications contrast two forms of bootstrap resampling in regression, illustrating their differences in a series of examples that include outliers and heteroscedasticity. Other regression examples use the bootstrap to estimate standard errors of robust estimators in regression and indirect effects in path models. Numerous variations of bootstrap confidence intervals exist, and examples stress the concepts that are common to the various approaches. Suggestions for computing bootstrap estimates appear throughout the discussion, and a section on computing suggests several broad guidelines.

An Introduction to Bootstrap Methods

Examples and Ideas

ROBERT STINE
University of Pennsylvania

The bootstrap is an *approach* to estimating sampling variances, confidence intervals, and other properties of statistics. Just as maximum likelihood refers to an estimation strategy rather than to any specific estimator, bootstrapping is a methodology for *evaluating* statistics based on an appealing paradigm. This paradigm arises from an analogy in which the observed data assume the role of an underlying population. As a result, bootstrap variances, distributions, and confidence intervals are obtained by drawing samples from the sample.

Data analysis seeks answers to questions such as "Does a new drug cure more people than the old one?" or "What factors affect how someone votes in an election?" Statistical answers to such questions require models that characterize the random behavior of observed factors. Estimates of the model arise from observed data and lead to description or inference. The importance of the bootstrap lies in this inferential step: The bootstrap gives standard errors and confidence

SOCIOLOGICAL METHODS AND RESEARCH, Vol. 18, Nos. 2 & 3,
November 1989/February 1990 243-291; © 1989 Sage Publications, Inc.

intervals that are typically better than alternatives that rely on untested assumptions. The flexibility of the bootstrap gives the data analyst the freedom to choose statistics whose standard errors would otherwise be difficult to measure. The bootstrap offers reliability and brings new insights to some of the difficult problems of data analysis.

Bootstrap calculations are typically computationally demanding. The bootstrap replaces difficult mathematics with an increase of several orders of magnitude in the computing needed for a statistical analysis. Rather than computing one or two sets of regression coefficients, bootstrapping easily entails several thousand. The computing demands of the bootstrap made such a strategy unthinkable until recently (Efron, 1979a). This trend toward greater use of computers can be expected to continue. As Tukey (1986) put it "In a world in which the price of calculation continues to decrease rapidly, but the price of theorem proving continues to hold steady or increase, elementary economics indicates that we ought to spend a larger and larger fraction of our time on calculation" (p. 74).

THE KEY IDEAS: BOOTSTRAPPING THE MEAN

The problem of estimating the variance of a sample mean illustrates the basic ideas. This friendly context permits the introduction of new topics without the added complexity of intricate statistical methods. The example also introduces some needed notation.

Probability distributions play a large role in the bootstrap. First let $X = (x_1, x_2, \ldots, x_n)$ denote a random sample of size n from the same population with mean μ and variance σ^2. If we let F denote the cumulative distribution function of the population, then $F(x) = Pr(x_i < x)$. In this notation, each x_i is a random variable having the cumulative distribution F, which is abbreviated $x_i \sim F$. Very often the population is assumed to be Gaussian (or normal), in which case F(x) is the function that appears in tables at the back of many statistics texts.

The sample-to-sample variation of the sample average is well known. If $\overline{x} = \Sigma\, x_i/n$ denotes the sample mean, then its variance is

$$VAR(\overline{x}) = \Sigma\ VAR\ (x_i)\ /\ n^2 = \sigma^2/\ n$$

When σ^2 is not known, the sample variance $s^2 = \Sigma(x_i - \overline{x})^2/(n-1)$ replaces it, giving the familiar estimator $var(\overline{x}) = s^2/n$. How well $var(\overline{x})$ estimates $VAR(\overline{x})$ depends on how close the distribution of the x_i is to being Gaussian with variance σ^2; the distribution does not need to drift far for

$\text{var}(\overline{x})$ to perform poorly.[1] Notice the notation: "VAR" written in upper case denotes the true variance, whereas "var" in lower case denotes an estimator of "VAR."

The bootstrap approach to estimating $\text{VAR}(\overline{x})$ is suggested by thinking about what $\text{var}(\overline{x})$ estimates: the variability of $\overline{x}$ across samples from the population with distribution F. In a *utopian setting* in which many samples from the population are available, formulas like that for $\text{var}(\overline{x})$ are unnecessary because one could compute the mean of many samples and estimate its variability directly. Many samples from the same population are seldom available, however, for a variety of reasons ranging from temporal changes to financial hurdles (see Finifter, 1972, which also refers to resampling as bootstrapping). Although it is not possible to get many samples from the population described by F, it is possible to get repeated samples from a population whose distribution approximates F. This is the idea behind the bootstrap: Replace the unknown function F, which describes a population that cannot be resampled, with an estimator of F, which describes a population that can be sampled repeatedly.

Given a minimum of assumptions, the optimal estimator of F is the *empirical distribution function* (EDF). For a sample of size n, the EDF is denoted F_n, and it is the cumulative distribution of the sample,

$$F_n\,(x) = \#(x_i \leq x)/n$$

where $\#(x_i \leq x)$ is the number of times that the inequality holds as i ranges from 1 to n. Thus, $F_n(10)$ is the proportion of the n sample observations that are less than or equal to 10. Unlike smooth distributions, such as the Gaussian, an empirical distribution has a "jump" at each observed value. For example, if the sample X consists of the five observations (1, 3, 4, 5.5, 8), then $F_5(0.5) = 0$, because none of the x_i are less than or equal to 0.5, and $F_5(5.5) = 4/5$. The jumps reflect the fact that only n distinct values are possible from this approximation to the true population F. Figure 1 shows the empirical distribution and the underlying population distribution for a sample of 25 Gaussian observations. Generally, F_n resembles F, but the jumps make it "rougher."

Given our willingness to approximate the population distribution F by the empirical distribution F_n two avenues are available for finding bootstrap variances. One is based on mathematics similar to those leading to $\text{var}(\overline{x})$, and the second relies on simulation. The latter computational approach to the bootstrap is a Monte Carlo simulation in

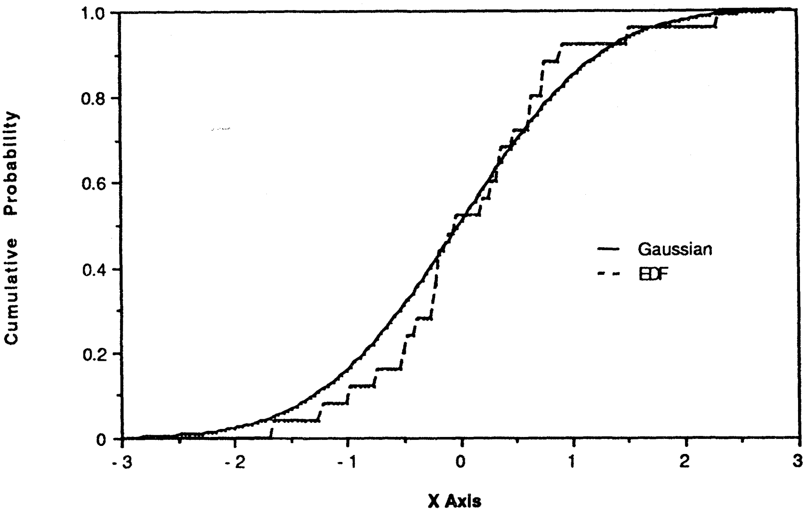

Figure 1 The empirical distribution of a sample of 25 observations tracks the cumulative Gaussian but is rougher.

which a multitude of samples are drawn from the observed data rather than from some hypothetical distribution. One draws repeated samples of the same size as the observed sample *with replacement* from the data, computes the mean of each such *bootstrap sample,* and then calculates the variance of this set of means. This variance estimate replaces var($\overline{x}$). Because each bootstrap sample consists of n observations that are drawn with replacement from the data, each bootstrap sample typically omits several observations and has multiple copies of others. If n is 5, for example, two bootstrap samples are (x_4, x_1, x_3, x_4, x_2) and (x_3, x_1, x_1, x_5, x_3). In the following algorithm, a superscript "*" distinguishes bootstrap quantities, as in var* and $\overline{x}^*$. The superscript b always ranges from 1 to B, indexing the bootstrap samples. The bootstrap algorithm for estimating the variance of $\overline{x}$ is:

(1) Use a random number generator to create *bootstrap samples of size n* by sampling *with replacement* from the observations. The *b*th bootstrap sample is denoted

$$X^{*(b)} = (x_1^{*(b)}, x_2^{*(b)}, \ldots, x_n^{*(b)})$$

where each $x_i^{*(b)}$ is a random selection from the original sample.

(2) Compute the mean $\bar{x}^{*(b)}$ from each of the bootstrap samples

$$\bar{x}^{*(b)} = \sum_{i=1}^{n} x_i^{*(b)}/n$$

(3) Use the B bootstrap means $\bar{x}^{*(1)}, \ldots, \bar{x}^{*(B)}$ to calculate the simulated *bootstrap variance estimate*

$$\text{var}_B^*(\bar{x}) = \sum_{b=1}^{B} \{\bar{x}^{*(b)} - \text{avg}(\bar{x}^{*(b)})\}^2/(B-1)$$

where $$\text{avg}(\bar{x}^{*(b)}) = \sum_{b=1}^{B} \bar{x}^{*(b)}/B$$

The number of bootstrap replications B depends on the application, but for standard error estimates, B ≈ 100 is generally sufficient. Because one can seldom draw every possible bootstrap sample, $\text{var}_B^*(\bar{x})$ estimates the bootstrap variance of the mean, $\text{VAR}^*(\bar{x})$, that we would get if B were infinitely large.[2]

In the setting of the sample mean, we can mathematically derive $\text{VAR}^*(\bar{x})$ without a computer simulation. Computer simulation is no more needed to find $\text{VAR}^*(\bar{x})$ than to determine $\text{VAR}(\bar{x})$. The mathematics are just like those leading to $\text{VAR}(\bar{x}) = \sigma^2/n$, once we accept replacing the theoretical distribution F with the empirical distribution F_n. In as much as the bootstrap samples are drawn with replacement from the observed sample, the empirical distribution function F_n defined by the original data *is* the cumulative distribution of the bootstrap samples. Rather than having samples in which $x_i \sim F$, we have bootstrap samples in which $x_i^* \sim F_n$. Thus, the population distribution is known for the bootstrap samples, and various theoretical calculations are possible. For example, the expectation of an observation from the population with distribution F is $E(x_i) = \mu$. Because F is not known, we do not know this expectation. On the other hand, let E^* denote expectation with respect to the bootstrap population that is defined by F_n. Because F_n is known, we can find expectations such as $E^*(x_i^*)$.

To evaluate $E^*(x_i^*)$, recall that the expectation of a random variable is a weighted average of the possible values, with weights given by the probability of that value occurring. Under bootstrap resampling from F_n, each of the observed values $x_i, \ldots, x_n$ is equally likely to occur with probability 1/n. Thus the expected value of a random draw from the population defined by F_n (which is just a random draw from the original observations) is the sample mean

$$E^*(x_i^*) = x_1(1/n) + x_2(1/n) + \ldots + x_n(1/n) = \bar{x}$$

Just as μ is the mean of the theoretic population, $\bar{x}$ is the mean of the bootstrap population. Because the variance is also defined as an expectation, it too is a weighted average

$$\begin{aligned} VAR^*(x_i^*) &= E^*\{x_i^* - E^*(x_i^*)\}^2 \\ &= (x_1-\bar{x})^2(1/n) + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2(1/n) \\ &= s_n^2 \end{aligned}$$

which is n/(n−1) times the usual variance estimator.[3] The x_i^* are drawn from the observations with replacement, so the x_i^* are independent of each other, and the bootstrap variance of the sample mean is

$$VAR^*(\bar{x}^*) = \Sigma\ VAR^*(x_i^*)/n^2 = s_n^2/n$$

which is almost the classical estimate of variance. For a large class of familiar statistics which includes certain regression models, simulation is not needed to obtain bootstrap variance estimates. (See the section on bootstrapping a regression model.)

Whether the bootstrap estimate is obtained by mathematics or simulation, the validity of bootstrap variance estimates requires that a key analogy holds for the statistic of interest. The key step of the bootstrap approach is to replace utopian sampling from the population defined by F with bootstrap resampling from the data, the population defined by F_n.[4] When is this a reasonable thing to do? Clearly, it depends on F_n being a good estimator of F. Without making other assumptions about the nature of the population, such as symmetry, F_n is about the best we can do.[5] The key analogy is that the *resampling* properties of $\bar{x}^*-\bar{x}$ must be similar to the *sampling* properties of $\bar{x}-\mu$. This analogy does not hold for every statistic. Because the bootstrap does not always give the correct answer, it is important to recognize the limits of this methodology, and some examples where it fails appear in the final section of this article.

TABLE 1
Summary of simulated lengths and coverages of 90% confidence intervals for the mean using 1000 samples of size 20 from a standard Gaussian population.

Method	Length Average	Length Std Dev	Coverage
Classical t	0.761	0.12	0.90
BS Percentile			
(B= 19)	0.795	0.20	0.87
(B= 99)	0.719	0.13	0.88
(B=499)	0.710	0.12	0.88

BOOTSTRAP DISTRIBUTIONS AND CONFIDENCE INTERVALS

The utility of a variance estimate depends upon how well we can use that estimate to measure the uncertainty in a statistic. For example, a common approximation uses the interval [estimator ± 2(standard error of estimator)] as a confidence interval. Under certain conditions – such as the statistic being unbiased with a symmetric distribution, e.g., the Gaussian – this interval approximates a 95% confidence interval. Whether we obtain the standard error via traditional methods or the bootstrap, the interval requires certain assumptions.

The bootstrap distribution of the statistic permits a more direct approach. The idea is to use percentiles of the *bootstrap distribution* of $\overline{x}^*$, $G^*(x) = Pr^*(\overline{x}^* \leq x)$ to determine a confidence interval. The 90% bootstrap percentile interval for μ is the interval that contains the middle 90% of the B bootstrap means. Symbolically, this interval is $[G^{*-1}(0.05), G^{*-1}(0.95)]$, where $G^{*-1}(p)$ denotes the *p*th quantile of the distribution of $\overline{x}^*$.[6] Whereas G^* is usually approximated by a simulation, we estimate it with $G_B^*(x) = \#\{\overline{x}^{*(b)} \leq x\}/B$. Again, B denotes the number of simulated bootstrap samples. The approximate bootstrap interval is then $[G_B^{*-1}(0.05), G_B^{*-1}(0.95)]$, the interval formed by the 5th and 95th percentiles of the B bootstrap means $\overline{x}^{*(1)}, \ldots, \overline{x}^{*(B)}$.

Table 1 compares the bootstrap 90% confidence intervals for the mean to the usual t-interval. The population is Gaussian so that the t interval $[\bar{x}+t(.05, n-1)\, s/\sqrt{n}, \bar{x}+t(.95, n-1)s/\sqrt{n}]$ is correct, where t(p,df) is the *p*th percentile of a t distribution with *df* degrees of freedom. The bootstrap interval is based on either B = 19, 99, or 499 bootstrap samples. The results in the table are from 1000 simulated samples of size 20 from a standard Gaussian distribution with $\mu = 0$ and $\sigma = 1$. The coverage column for each interval is the proportion of the 1000 intervals that included the true mean 0. Even with only 19 bootstrap samples, the bootstrap intervals nearly obtain the performance of the best interval in this case, that given by *knowing* that the data are from a Gaussian distribution. Because the standard error of the coverage estimates is roughly $\{(.1)(.9)/1000\}^{1/2} \approx 0.01$, the percentile intervals have coverage significantly less than 0.9. Increasing the number of bootstrap replications B does little to improve the coverage, although it does lead to a shorter interval whose length is more stable from sample to sample. The slight lack in coverage and more variable length are the price we pay for not assuming a Gaussian distribution — a small cost given that data are seldom from a Gaussian distribution.

RELATIONSHIP TO THE JACKKNIFE

Sample re-use methods such as the bootstrap are not entirely new, and perhaps the most well-known predecessor is the jackknife. The jackknife shares the goal of easily obtained, trustworthy variance estimates, but it relies on a less demanding computational algorithm. Rather than compute the statistic for a large collection of bootstrap samples from the original data, the jackknife relies on dividing the sample observations into, say, S disjoint subsets, each having the same number of observations. The statistic of interest is then computed S times, each time omitting one of the subsets. Rather than having perhaps 100 repetitions of the statistic, the jackknife requires at most n when each subset consists of a single observation. Many overviews of the jackknife exist, such as that of Miller (1974) and the applications of Mosteller and Tukey (1977). A recent study of the theoretical properties of the jackknife and related methods appears in Wu (1986).

To illustrate the jackknife, consider again the problem of estimating the variance of $\bar{x}$. For this example, let each subset consist of one observation. Begin by computing the mean of $(x_2, x_3, \ldots, x_n)$, the n−1 observations left after removing x_1. Label the mean of these n−1 observations $\bar{x}_{(-1)}$. Then compute $\bar{x}_{(-2)}$, the mean of $(x_1, x_3, \ldots, x_n)$. Continuing in this fashion, the procedure leads to n "leave-one-out" means $\bar{x}_{(-1)}, \ldots, \bar{x}_{(-n)}$. The jackknife combines these to obtain its variance estimate. Unlike the bootstrap values $\bar{x}^{*(b)}$, which are independent of each other (conditional on the observed sample), the jackknife replicates $\bar{x}_{(-i)}$ are highly correlated; every pair of jackknife means has n-2 observations in common. By comparison, given the values in the sample, the bootstrap replicates $\bar{x}^*$ are conditionally independent of each other; two bootstrap samples may have no values in common. A further difference from the bootstrap lies in the sample size. The jackknife "samples" are of size n–1 rather than n. As a result, the jackknife variance expression includes an adjustment factor of (1–1/n). The jackknife variance estimate is

$$\mathrm{var}_{JK}(\bar{x}) = \frac{n-1}{n} \sum_{i=1}^{n} (\bar{x}_{(-i)} - \bar{x}_{(\cdot)})^2$$

where $\bar{x}_{(\cdot)}$ is the average of $\bar{x}_{(-1)}, \ldots, \bar{x}_{(-n)}$.

Efron (1982) presents the bootstrap as a generalized framework encompassing the jackknife. For example, he shows that the jackknife estimate of variation is an approximation to the bootstrap estimate and agrees with the bootstrap in large samples for statistics such as the mean. Although such an embedding is possible, the jackknife takes a fundamentally different view of the possible replicates of the statistic. The jackknife treats them as a finite collection, whereas bootstrap resampling assumes that the replicates are a sample from a population of infinite size. As a result, the jackknife only uses at most n values of the statistic, whereas the bootstrap considers a much larger collection.

The degree to which the bootstrap outperforms the jackknife (or vice versa) seems to depend on the degree to which the data are really independent observations from the same population. If some complex structure or correlation exists among the observations, a grouped jackknife may be more appropriate (Tukey, 1987). Notice also that the jackknife was never intended to be a method for estimating a distribution. The usual procedure for getting confidence intervals from the jackknife is to use the jackknife standard error in a modified t-interval, as described in the initial abstract of Tukey (1958).

THE ROLE OF MATHEMATICS

Most of the mathematical results about the bootstrap describe how it performs as the sample size grows. For example, one can show that the bootstrap variance estimates in regression approach the correct value as the sample size becomes arbitrarily large (Freedman, 1981). *Asymptotic*, or large sample, properties of a statistic are important theoretical results, but they are not always indicative of the performance of a statistic in applications. No one wants to use an estimator that is not consistent, and asymptotics are needed to determine such large sample properties. For data analysis, however, asymptotics are a guide that can often be unreliable when confronted with small samples and outlying values. Most asymptotic results describe how a statistic behaves under the best of circumstances: unlimited growth of the number of independent observations from the same population.

These assumptions suggest two questions to ask when assessing the relevance of asymptotics. First, How large a sample size is needed for the asymptotics to be useful? Some asymptotic results, such as the familiar Central Limit Theorem, are relevant even when the sample has only 30 observations. In fact, until recently, a commonly used method of using computers to generate samples from a normal distribution was to take the average of 12 uniformly distributed observations. Other asymptotic results require larger samples. For example, the asymptotics that lead to the standard error estimates in the section on applications in robust regression require larger samples than the 20 observations used there. The second question asks: Can a large sample really consist of independent observations from the same population? Such an abstraction is needed for the mathematics, but is unusual in practice. If it requires six months to gather the data for a large survey, then the underlying population may have changed over the course of the data collection (see Cook and Campbell, 1976).

OVERVIEW OF THE REMAINING SECTIONS

The following section illustrates the bootstrap in regression and constitutes the bulk of this article. The section on bootstrap confidence intervals describes bootstrap confidence intervals in detail. The fourth section, on computing bootstrap estimates, contains advice, including deciding on how large to set the number of replications B. Finally, the fifth section suggests where resampling methodology is moving, including some adventurous applications. Other reviews of the bootstrap tend to focus on the question, "Does the bootstrap work?" (e.g., Diaconis and Efron, 1983; Efron, 1979b; Efron, 1982; Efron and Gong, 1983; Efron and Tibshirani, 1986). With some exceptions, this review begins with the premise that the bootstrap is a good idea and focuses on how to use it in data analysis, particularly regression.

APPLICATIONS OF BOOTSTRAP RESAMPLING IN REGRESSION

Regression models remain at the heart of applied statistics. Robust estimation strategies and residual diagnostics have improved the usefulness of these techniques, and the bootstrap adds another dimension. After a quick review of the basic regression model, we describe two methods of bootstrap resampling. The nature of the data determines which alternative is appropriate. Modeling assumptions are very important with the regression model, and heteroscedasticity and serial correlation present problems that the bootstrap, if properly used, often handles better than classical methods. Once the model is chosen, the bootstrap also allows us to compare different estimation strategies, such as robust regression estimates to least squares. Some problems in structural equations also are easily handled with the bootstrap.

BOOTSTRAPPING A REGRESSION MODEL

In the usual regression model, $\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$ denotes the n×1 vector of the response, and the n×k matrix of regressors is $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)'$, where the k×1 vector $\mathbf{x}_i$ denotes the regressors for the *i*th observation. The usual linear model is then

$$\mathbf{Y} = \mathbf{X}\,\beta + \varepsilon \text{ or } y_i = x_i'\beta + \varepsilon_i,\ i = 1, \ldots, n \qquad [1]$$

where ε is an n×1 vector of uncorrelated error terms having mean 0 and variance σ^2. The k×1 vector β holds the unknown parameters, for which the ordinary least squares (OLS) estimator is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \qquad [2]$$

It follows that $\mathrm{VAR}(\hat{\beta}) = \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}$. Because σ^2 is not usually known, $\mathrm{VAR}(\,\hat{\beta}\,)$ is estimated by

$$\mathrm{var}(\hat{\beta}) = s^2(\mathbf{X}'\mathbf{X})^{-1} \qquad [3]$$

where s^2 is the unbiased variance estimator provided by the residuals $e_i = y_i - x_i'\,\hat{\beta}$, $i = 1, \ldots, n$

$$s^2 = \Sigma\,(y_i - x_i'\,\hat{\beta}\,)^2/(n-k) = \Sigma\,e_i^2/(n-k).$$

Throughout this section, s^2 denotes this variance estimator, not the sample variance estimator of the introduction. A general reference on the theory of least squares estimation is Fox (1984).

Two methods exist for bootstrapping the regression model, and the choice of which to use depends upon the regressors. If the regressors are *fixed*, as in a designed experiment, then bootstrap resampling must preserve that structure. Each bootstrap sample should have the same regressors. On the other hand, regression models built from survey data typically have regressors that are as random as the response, and bootstrap samples should also possess this additional variation.

Resampling with random regressors. Bootstrapping regression models with *random* regressors follows the strategy of the introduction. Let the (k+1)×1 vector $\mathbf{z}_i = (y_i, \mathbf{x}_i')'$ denote the values associated with the *i*th observation. Just as in the case for bootstrapping $\bar{x}$, one samples with replacement from the observations. Only in this case, the set of observations are the vectors $(\mathbf{z}_i, \ldots, \mathbf{z}_n)$ rather than a set of scalar values. The three steps of the random regressor algorithm are:

(1r) Draw a bootstrap sample $(\mathbf{z}_1^{*(b)}, \mathbf{z}_2^{*(b)}, \ldots \mathbf{z}_n^{*(b)})$ from the observations and label the elements of each vector

$$\mathbf{z}_i^{*(b)'} = (y_i^{*(b)}, \mathbf{x}_i^{*(b)'})'$$

From these form the vector $\mathbf{Y}^{*(b)} = (y_1^{*(b)}, y_2^{*(b)}, \ldots, y_n^{*(b)})'$ and the matrix $\mathbf{X}^{*(b)} = (\mathbf{x}_1^{*(b)}, \mathbf{x}_2^{*(b)}, \ldots, \mathbf{x}_n^{*(b)})'$.

(2r) Calculate the OLS coefficients from the bootstrap sample:

$$\hat{\beta}^{*(b)} = (\mathbf{X}^{*(b)'}\mathbf{X}^{*(b)})^{-1}\, \mathbf{X}^{*(b)'}\mathbf{Y}^{*(b)}$$

In a very small sample, the matrix $(\mathbf{X}^{*(b)'}\mathbf{X}^{*(b)})$ might be singular and would necessitate drawing a new sample.

(3r) Repeat steps 1 and 2 for b = 1, . . . , B, and use the resulting bootstrap estimates $\hat{\beta}^{*(1)}, \hat{\beta}^{*(2)}, \ldots, \hat{\beta}^{*(B)}$ to estimate variances or confidence intervals. The bootstrap estimate of the covariance matrix of $\hat{\beta}$ is

$$\text{var}_B{}^*(\hat{\beta}^*) = \sum_{b=1}^{B} \mathbf{d}_b{}^*\mathbf{d}_b{}^{*\prime}/(B-1) \qquad [4]$$

where the vector of deviations is $\mathbf{d}_b{}^* = \hat{\beta}^{*(b)} - \text{avg}(\hat{\beta}^{*(b)})$, b = 1,. . . ,B.

The effect of sampling the $\mathbf{z}_i$ is to keep the response y_i of a given observation paired with the regressors $\mathbf{x}_i$ of that observation.[7]

Resampling with fixed regressors. When the regression model has fixed regressors, the resampling should preserve the structure of the design matrix. For example, suppose **X** defines a balanced two-way analysis of variance for an experiment or consists of polynomial time trends. If random resampling is used, $\mathbf{X}^*$ would not likely possess the needed structure: The ANOVA design would be unbalanced and the polynomial time trends would have gaps and clusters. In regression models in which **X** is fixed, each utopian sample has the same design. Bootstrap samples need this same characteristic. The algorithm that produces this behavior is not as straightforward as that for the random design and relies on regression residuals. It is, however, more computationally efficient. The change in the preceding algorithm occurs in the first step, which defines how the bootstrap samples are generated. The second step reveals the computational advantage. The steps are:

(1f) Compute the bootstrap samples by adding *resampled residuals* onto the least squares regression fit, holding the regression design fixed:

$$\mathbf{Y}^{*(b)} = \mathbf{X}\,\hat{\beta} + \mathbf{e}^{*(b)}$$

where the n×1 vector $\mathbf{e}^{*(b)} = (\mathbf{e}_1^{*(b)}, \mathbf{e}_2^{*(b)}, \ldots, \mathbf{e}_n^{*(b)})'$, and each $\mathbf{e}_i^{*(b)}$ is a random draw from the set of n regression residuals.

(2f) Obtain least squares estimates from the bootstrap sample:

$$\begin{aligned}\hat{\beta}^{*(b)} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}^{*(b)}\\ &= \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}^{*(b)}\end{aligned}$$

Because $(\mathbf{X}'\mathbf{X})^{-1}$ appears in every $\hat{\beta}^{*(b)}$, only one matrix inverse is needed. Also, the second line shows that one need never explicitly form $\mathbf{Y}^*$.

(3f) Repeat steps (1) and (2) for b = 1,. . . , B, and proceed as in (3r).

In contrast to the random regressor model, this resampling approach generates $\mathbf{Y}^*$ by adding samples of the residuals to the fitted equation $\mathbf{X}\,\hat{\beta}$ rather than by resampling from the actual data.

The introduction of residuals raises important issues. Residuals are the product of a model imposed upon the data; their values depend upon the model that we choose. Ideally, one would like to be able to sample from the true population of errors. Because this population is unknown, the bootstrap resamples the residuals, even though least squares residuals are rather different from the true errors. In particular, residuals are neither independent nor identically distributed, even if the regression model is correct. The covariance matrix of the residual vector **e** is

$$\mathrm{VAR}(\mathbf{e}) = \mathrm{VAR}\{(\mathbf{I}-\mathbf{H})\varepsilon\} = \sigma^2(\mathbf{I}-\mathbf{H})$$

where the **I** is the nxn identity matrix and the projection or "hat" matrix **H** is

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \qquad [5]$$

Typically, **I–H** has substantial off-diagonal elements and a non-constant diagonal. Hence, in place of independent, constant variance errors, the bootstrap samples correlated heteroscedastic residuals.[8] The bootstrap succeeds in spite of these differences between residuals and true errors. The bootstrap vector $\mathbf{e}^*$ consists of independent errors with constant variance, *regardless* of the properties of the residuals. Bootstrap resampling from the residuals gives independent error terms with variance $(1-k/n)\,s^2$.[9] Inasmuch as the variance of the population defined by the residuals is too small, one can "fatten" the residuals by dividing each by a factor of $(1-k/n)^{1/2}$. Tukey (1987) heartily recommends such "degree-of-freedom" corrections, and these modifications are useful in contexts such as prediction intervals (Stine, 1985).

As in the case of the sample mean, we do not need simulation to find the bootstrap variance of $\hat{\beta}$ in the fixed regressor model. The mathematical derivation of the bootstrap variance mimics the usual derivation of VAR($\hat{\beta}$). For models with fixed regressors, the bootstrap variance is (Efron, 1982: 36)

$$\text{VAR}^*(\hat{\beta}^*) = (1-k/n)\,s^2\,(\mathbf{X}'\mathbf{X})^{-1}$$

differing only by a scale factor from var($\hat{\beta}$). In fact, for any *linear statistic*, one can compute the bootstrap variance without computer simulation.[10]

EFFECTS OF OUTLIERS

So how do these resampling methods compare, how are they useful in data analysis, and which is better to use? The choice of which to use is easy to determine: Bootstrap resampling should always resemble the original sampling procedure. Models having a fixed design lead naturally to bootstrapping with fixed regressors; those with a random design lead to random resampling. Although these two methods of resampling differ, it has been proven that the resulting differences become small as the sample size grows (Freedman, 1981). But with small samples, important differences emerge that must be understood if one is to make practical use of the bootstrap. In particular, because these methods differ in how they bind y_i to x_i, they react differently to outliers.

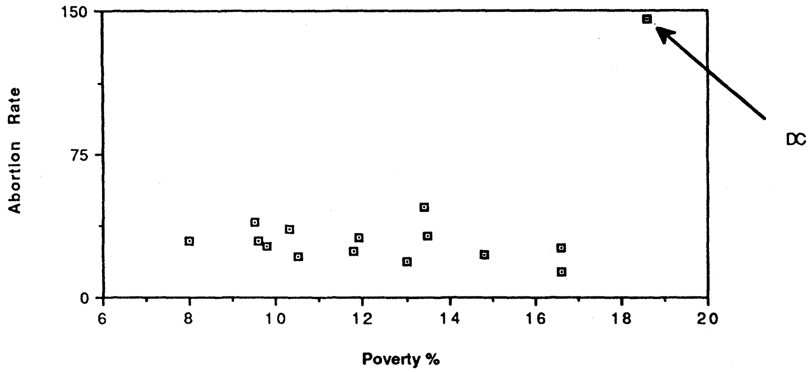

Figure 2 Abortion rates and poverty in the East, with the District of Columbia being a large outlier.

Some ideas from regression diagnostics are needed for this discussion. The *leverage* of an observation in a regression model is a measure of how sensitive $\hat{\beta}$ is to changes in the response at that point. The leverage values H_i, i = 1,. . . ,n are the diagonals of the projection matrix **H** (5) and $0 \leq h_i \leq 1$. A related measure is the *influence* of an observation, which indicates how much $\hat{\beta}$ changes when an observation is removed from the data. Influence combines the leverage and residual for an observation; the change in $\hat{\beta}$ when the ith observation is removed is $(X'X)^{-1}\ x_i e_i\ /(1-h_i)$ (e.g., Fox, 1984).

An example illustrates the different effects of outliers. The scatterplot of Figure 2 shows the abortion rate per 1000 women aged 18-44 versus the proportion of the population below the poverty level in the 14 Eastern coastal states as well as the District of Columbia. (The data are from Tables 104 and 712 of *The Statistical Abstract of the U.S., 1988.*) The outlier is the District of Columbia, which had an abortion rate of approximately 146 per 1000 women in 1985. This observation is very influential in a simple regression of the abortion rate on the poverty percentage because it combines a large residual with high leverage. The high leverage is the result of D.C. having the largest poverty percentage (18.6%). Least-squares and robust estimates for a simple linear model appear in Table 2. The least-squares slope is positive if the outlier is included, suggesting to the hasty data analyst that higher poverty leads to higher abortion rates. Dropping D.C. suggests the opposite conclusion, although not very strongly. The robust fit resembles the least-squares fit that omits D.C.

TABLE 2
Estimated least squares and robust estimates with t-values for a linear regression model fit to the abortion rate data in Figure 2

Estimation Method	Intercept	Slope
Least squares, all data	-21.4 (-0.7)	+4.6 (1.8)
Least squares, excluding DC	+41.8 (3.8)	-1.1 (-1.3)
Biweight, all data	+42.3 (4.5)	-1.2 (-1.7)

The two regression resampling methods differ considerably when applied to these data. Because the poverty percentage in these states is random and not experimentally controlled, random resampling is the correct method. When it is used, D.C. appears in about 64% of the bootstrap samples.[11] Because samples containing the District of Columbia give a positive slope and those without this observation usually give a negative slope, the estimates of the slope from bootstrap samples are sometimes positive and sometimes negative. The histogram of the bootstrap slopes based on B = 500 bootstrap samples in Figure 3 shows the resulting bimodal shape, with most of the slopes being positive.[12] In contrast, the method of fixed regressors samples the residuals and adds them to the least-squares regression fit. Thus, the large residual for D.C. could appear at any (or even several) of the 15 observations in a bootstrap sample. By severing the tie between the large residual and high leverage point, residual resampling produces slope estimates whose histogram (also in Figure 3) is quite normal in appearance.

Neither bootstrap scheme "cures" the outlier problem, and they give different impressions of the effect of an outlier. Residual resampling spreads the effect of the outlier about the design, whereas random resampling keeps it localized as in the observed sample. Because the regressor in this problem is random, random resampling is appropriate. So then how are we to react to the bimodal shape of the histogram for the slope? It draws our attention to a problem in the regression model, and reveals how outliers affect the distribution of $\hat{\beta}$.[13]

BOOTSTRAPPING WITH HETEROSCEDASTICITY

Because an outlier can be viewed as an observation with large variance, the preceding example suggests that the two resampling methods react differently in the presence of heteroscedasticity. If the

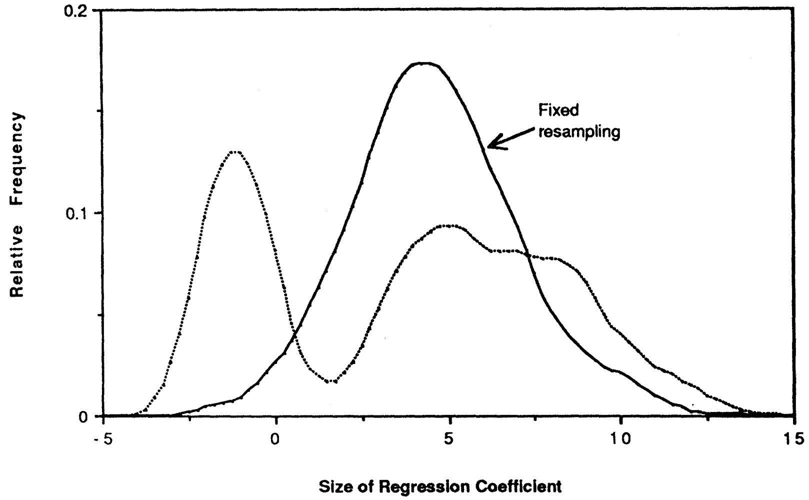

Figure 3: Bootstrap distributions of the least squares slope via fixed and random resampling.

errors are heteroscedastic and fixed resampling is used, random resampling of the residuals leads to bootstrap samples that are homoscedastic. Taking random draws from the residuals scatters the residuals around the design, giving bootstrap data sets that show no sign of heteroscedasticity. Random resampling of the observations preserves the heteroscedasticity. If VAR(ε) = $\sigma^2\mathbf{D}$, where $\mathbf{D}$ is an nxn diagonal matrix with varying entries that reflect the presence of heteroscedasticity, then residual resampling leads to the variance estimate

$$\text{VAR}^*_{\text{fixed}}(\hat{\beta}) = v^2(X'X)^{-1}$$

where $nv^2 = E(e'e) = \sigma^2$ trace $(\mathbf{I}-\mathbf{H})\mathbf{D}$. Random resampling preserves the heteroscedasticity, and its bootstrap variance is approximately the correct answer

$$\text{VAR}^*_{\text{random}}(\hat{\beta}) \approx \sigma^2(X'\ X)^{-1}(X'\mathbf{D}X)\ (X'X)^{-1}$$

An extensive discussion of the effects of uncorrected heteroscedasticity on the bootstrap appears in Wu (1986, especially the discussion).

Now suppose that we have recognized the presence of heteroscedastic errors and want to do something about it. The broad validity of random resampling makes it preferable to the usual approach of estimating the variance of weighted least squares (WLS) estimators, assuming that the regressors of the bootstrap sample possess the sampling characteristics of those in the original data. If we continue with the assumption VAR(ε) = $\sigma^2\mathbf{D}$, then the optimal WLS estimator and its variance are (e.g., Fox 1984)

$$\hat{\beta}_W = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{Y} \qquad [6]$$

$$\mathrm{VAR}\ (\hat{\beta}_W) = \sigma^2(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}$$

In practice, however, one often uses estimators such as

$$\hat{\mathbf{b}}_W = (\mathbf{X}'\hat{\mathbf{D}}^{-1}\mathbf{X})^{-1}\ \mathbf{X}'\hat{\mathbf{D}}^{-1}\mathbf{Y}$$

for which the obvious variance estimator corresponding to (6) is

$$\mathrm{var}\ (\hat{\mathbf{b}}_W) = s^2\ (\mathbf{X}'\hat{\mathbf{D}}^{-1}\mathbf{X})^{-1} \qquad [7]$$

where $\hat{\mathbf{D}}$ is estimated from the data. Thus $\hat{\beta}_W$ is the WLS estimator that requires that we know the variance structure in the matrix **D**; $\hat{\mathbf{b}}_W$ is a practical approximation to $\hat{\beta}_W$ based on an estimate of **D**. Unfortunately, the common variance estimator (7) is more an estimator of the variance of $\hat{\beta}_W$ than of VAR($\hat{\mathbf{b}}_W$) because it does not incorporate the estimation of **D**, which is needed in $\hat{\mathbf{b}}_W$.

Bootstrap methods capture more of the uncertainty induced by the estimation of error variance structure. As an example, suppose that our data have several observations at each value of the regressor, and the variance increases with the regressor. The book-price data (from Table 369 of the *Statistical Abstract of the U.S., 1988*) in Figure 4 have this form. The average and variance of price increase over time. One model for these data is to assume that $\mathbf{D} = \mathrm{diag}\ (d_1, d_1, d_1, d_1, d_1, d_1, d_2, d_2, \ldots, d_4)$. Because the error variances depend in some unknown fashion upon year, an iterative estimation strategy is needed:

(1) Use OLS to obtain residual estimates, $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\beta}$

(2) Estimate the variance of the residuals at each value of the regressor, and estimate **D** with

$$\hat{d}_i = \sum_{j=1}^{6} (e_{ij} - \bar{e}_i)^2/5$$

where e_{ij} is the *j*th residual at the *i*th design point and $\bar{e}_i$ is the average of the six residuals in that group.

(3) Form the estimator $\hat{\mathbf{b}}_W = (\mathbf{X}'\hat{\mathbf{D}}^{-1}\mathbf{X})^{-1}\ \mathbf{X}'\hat{\mathbf{D}}^{-1}\mathbf{Y}$.

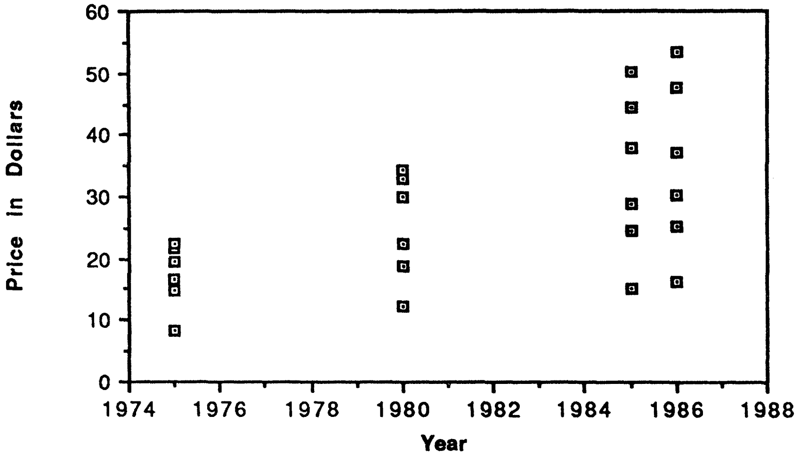

Figure 4: Bootstrap distributions of the least squares slope via fixed and random resampling.

(4) If the difference of $\hat{b}_W$ from the previous estimate of β is small, terminate the calculations. Otherwise, continue with step 5.

(5) Compute the residuals associated with the most recent estimate of β, $e = Y - X\hat{b}_W$, and return to step 2.

Each iteration attempts to obtain more accurate error estimates and use these to get a better idea of the error covariances. Applying this procedure to the book-price data gives the slope estimate 1.62. The nominal WLS variance estimator $s^2(X'\hat{D}^{-1}X)^{-1}$ on the first line of Table 3 ignores the estimation of **D**, and one suspects that this estimate is too small. By comparison, the bootstrap variance on line two of the table is 23% larger. Because the bootstrap estimate of variance incorporates the estimation of **D**, it seems to be a better estimate of variation than that from the traditional procedure. This bootstrap variance is obtained by repeating the preceding algorithm on 20 bootstrap samples using the following variant of random resampling. Inasmuch as errors in different years have different variances, we resample within each group. Each bootstrap sample thus has six observations at each of four years, preserving this structural feature of the original data.

It remains to be determined whether either of these variance estimates is accurate, and the bootstrap is one of three approaches to

TABLE 3
Standard Error estimates for the slope of the regression line fit to the heteroscedastic book-price data of figure 4.

Method of Estimation	Standard Error
(1) WLS estimate $s^2 (X'\hat{D}^{-1}X)^{-1}$	0.39
(2) Bootstrap, B=20	0.48
(3) Iterated WLS	0.36
(4) Iterated bootstrap	0.47

answering this question. Mathematical expressions would be best, but the iterative nature of $\hat{b}_W$ suggests that this approach is unlikely to yield easily interpreted results without a host of assumptions. Alternatively, we could perform a simulation in which we estimate the variation of $\hat{b}_W$ across simulated samples, and compare this variation to the average of the nominal variance estimates (7). However, we have to decide what distribution to sample. But the bootstrap is a procedure for evaluating statistics, and the bootstrap variance estimate is a statistic. So why not bootstrap the bootstrap? The amount of calculation becomes intimidating, but the strategy of using the bootstrap to evaluate itself is appealing and avoids the troublesome choice of what distribution to sample. Keep in mind that the bootstrap estimate of the variance of $\hat{b}_W$ is a statistic like any other, although it takes a bit more calculation to obtain. The nested computations proceed as follows:

(1) Draw B_1 bootstrap samples $(Y^*, X^*)^{(j)}$, $j = 1, \ldots, B_1$, from the original data. For each sample, estimate the variance of $\hat{b}_W$ using the WLS expression (7). Denote these estimates $\mathrm{var}_{wls}{}^{(j)}$, $j=1, \ldots, B_1$.

(2) For each of the B_1 initial bootstrap samples:

(2a) Draw B_2 bootstrap samples from $(Y^*, X^*)^{(j)}$ and label these $(Y^{**}, X^{**})^{(jb)}$, $b = 1, \ldots, B_2$.

(2b) Estimate $\hat{b}_w{}^{**(jb)}$ from each of the B_2 samples $(Y^{**}, X^{**})^{(jb)}$.

Compute the variance of the collection $\hat{b}_W{}^{**(jb)}$ as in (4), and denote this variance estimate $\mathrm{var}_{B2}{}^{*(j)}$.

(3) Compare the average WLS estimate to the average bootstrap estimate:

$$\text{average nominal WLS std. error: } \{\Sigma_j \mathrm{var}_{wls}{}^{(j)}/B_1\}^{1/2},$$

$$\text{average bootstrap std. error: } \{\Sigma_j \mathrm{var}_{B2}{}^{*(j)}/B_1\}^{1/2}.$$

Because we are sampling the population defined by the observed sample, the correct answer *is* the original bootstrap estimate given on line 2 of Table 3; that is, 0.48 *is* the standard error of the slope estimate when sampling from the population defined by the original data. The results of step 3 are on lines 3 and 4 of Table 3 with $B_1 = 100$ and $B_2 = 20$. The average 0.36 of the WLS estimates is too small; the estimation of **D** adds a substantial amount to the variance of the slope estimator. On the other hand, the average of the bootstrap samples 0.47 is quite close to the target value, 0.48.[14]

APPLICATIONS IN ROBUST REGRESSION

The notion of using the bootstrap to estimate variances of iterative estimators such as $\hat{b}_W$ suggests applications in robust regression. Bootstrap methods share the intent of robust statistics, albeit with a different slant. Robust statistical methods provide nearly optimal parameter estimates under a variety of broad conditions. Bootstrap methods reduce the need for tenuous assumptions, but concentrate on evaluating an estimator rather than defining it. Bootstrapping does not produce robust estimators, but it can suggest how robust an estimator is. For a well-motivated introduction and overview of robust methods, see Hampel et al. (1986).

In robust regression, outlying observations are downweighted so that the regression captures the pattern in the majority of data rather than tracks outliers. The downweighting is accomplished by using an iterative reweighting scheme not too different from that used to obtain $\hat{b}_W$. The weights are chosen by a variety of schemes, such as the *biweight* used here. Asymptotic methods exist for estimating the variances of robust regression coefficients, and one must take care to allow for the variability in the weighting process (Street et al., 1988). The bootstrap is also a valid procedure for estimating the variance of common robust estimators (Shorack, 1982).

Our example of bootstrapping a robust regression utilizes a quadratic model for the growth of the U.S. population, which appears in Figure 5. Despite a high R^2, the last few observations do not fit this model well and have relatively large, highly leveraged residuals. (This deviation from the model suggests the presence of specification error: A variable that is not included in this simple model assumes an important role in the later data.) These later points lead to the difference shown in Table 4 between the coefficients of the OLS fit and a robust fit. In both quadratic

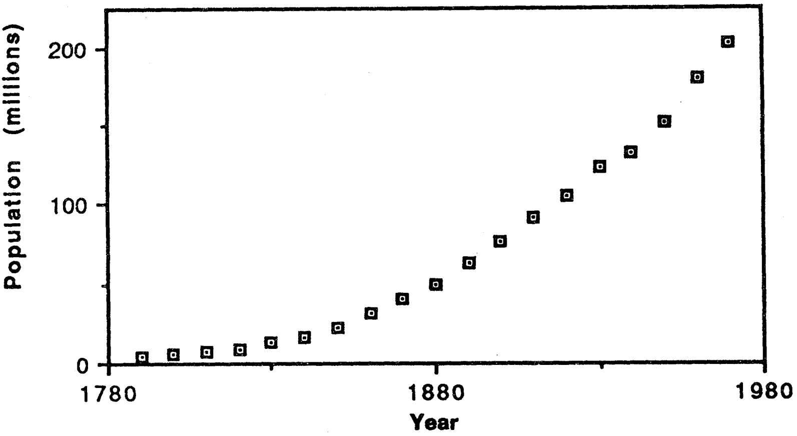

Figure 5: US Population shows quadratic growth.

models, the regressor (time) is centered and rescaled to run from −1 to 1. The slope estimates are similar, but the differences are large relative to the estimated standard errors. The difference between the coefficients of the linear term in the model is 1.8, which is more than four standard errors of the robust estimator. Also, the robust estimator claims a standard error that is less than half that of the least-squares estimator.

The bootstrap gives a different sense of how these estimators behave. We must use fixed resampling in this model because the regressors are time and time squared. Random resampling would yield bootstrap data sets with several observations at one year, and none at others. Fixed resampling preserves the rigid time progression of the original data and is the appropriate method. The average and standard error of the bootstrap results with B = 500 are in Table 5. The bootstrap results suggest that both coefficient estimators are unbiased for their expectations because the averages of the bootstrap coefficients approximate the original estimates. The bootstrap estimates of standard error are about the same as the usual OLS estimates. In fixed resampling, the bootstrap estimates of standard error of the coefficient estimators approach the usual least-squares values as B grows large. The bootstrap estimates of standard error in Table 5, however, are much larger than asymptotic standard error estimates for the robust coefficients in Table 4; the sample size is perhaps too small for the asymptotics to be accurate. In both cases, however, the robust estimator has smaller variance than the

TABLE 4

Coefficients of quadratic models for the U.S. population growth shown in Figure 5, estimated by least squares and robust methods.

	Coefficient		
Estimator	Constant	Linear	Quadratic
OLS	50.7 (0.96)	97.1 (1.05)	51.4 (1.93)
Robust	51.1 (0.37)	98.9 (0.40)	52.8 (0.74)

OLS estimator.[15] The bootstrap distributions of the OLS and robust linear coefficients in Figure 6 confirm these impressions: The distribution of the robust estimator is much more tightly concentrated than that of the OLS estimator.

A closer look at the bootstrap distribution of the robust estimator also suggests why its bootstrap standard error is so much larger than the asymptotic estimate. The bootstrap distribution of the robust estimator in Figure 6 is very peaked, but this figure conceals the tails of the distribution. The Gaussian quantile comparison plot in Figure 7 reveals that the tails of this distribution are much heavier than those of the Gaussian distribution. The asymptotic standard error is based on a Gaussian approximation to a sampling distribution. Because the robust estimator has a long-tailed distribution, this approximation is not very accurate. It is ironic that one typically sees such Gaussian properties as standard error applied to estimators like the biweight that do not show Gaussian behavior in small samples, an observation I owe to John Fox. A better comparison of these estimators would be in terms of a more robust estimator of variation, such as the hinge-spread. On the other hand, once standard errors are surrendered, it becomes quite hard to draw comparisons to traditional results.

STRUCTURAL EQUATION MODELS AND THE BOOTSTRAP

Some of the more interesting features of path models are indirect effects. These measure the effect that one variable has upon another through other factors in the model. Because estimates of indirect effects are products of several regression coefficients, one cannot apply the usual least-squares formulas. Some applications of indirect effects appear in Fox (1984).

TABLE 5

Average and standard error of bootstrap replicates of the least squares and robust coefficients in models for U.S. population growth (B=5000).

	Coefficient		
Estimator	Constant	Linear	Quadratic
OLS	50.7 (0.90)	97.0 (0.97)	51.4 (1.79)
Robust	51.1 (0.58)	98.9 (0.73)	52.6 (1.33)

One approach to estimating the distribution of an estimated indirect effects is to combine the *delta method* with a normal approximation. The delta method (Bishop et al., 1975) is based on the observation that it is easy to compute the variance of linear functions of statistics, such as $a+b\hat{\theta}$. The idea behind the delta method is to find a linear approximation to a statistic, and use this approximation to estimate the variance. Suppose $g(\hat{\theta})$ is a nonlinear function of the statistic $\hat{\theta}$, for example, $g(\hat{\theta}) = \log(\hat{\theta})$. To approximate the variance of $g(\hat{\theta})$, we estimate the variance of the linear approximation based on the derivative of g at $\hat{\theta}$. This approximation to $g(\hat{\theta})$ is $g(\theta)$ plus a slight change to reflect replacing θ by $\hat{\theta}$. The change in g as the parameter changes from θ to $\hat{\theta}$ is estimated by the derivative of g at θ, $g'(\theta)$. The resulting approximation to $g(\hat{\theta})$ is then

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta}-\theta) \qquad [8]$$

For $g(\hat{\theta}) = \log(\hat{\theta})$, this approximation is $\log(\hat{\theta}) \approx \log(\theta) + (\hat{\theta}-\theta)/\theta$ because $g'(\theta = 1/\theta$. In as much as $g(\theta)$ and $g'(\theta)$ are constants, the variance of $g(\hat{\theta})$ is approximately

$$\text{VAR}\ \{g(\hat{\theta})\} \approx g'(\theta)^2\ \text{VAR}(\hat{\theta}) \qquad [9]$$

In practice, (9) is not useful because it requires evaluating the derivative at the true parameter and finding $\text{VAR}(\hat{\theta})$, both of which are usually unknown. Substituting $\hat{\theta}$ for θ and $\text{var}(\hat{\theta})$ for $\text{VAR}(\hat{\theta})$ in (9), one is led to the delta method variance approximation

$$\text{var}\{g(\hat{\theta})\} \approx g'(\hat{\theta})^2\ \text{var}(\hat{\theta}) \qquad [10]$$

The accuracy of $\hat{\text{var}}\{g(\hat{\theta})\}$ depends on three factors: the distance of $\hat{\theta}$ from θ, the smoothness of the derivative that permits the switch from $g'(\theta)$ to $g'(\hat{\theta})$, and the accuracy of the initial approximation (8). In the

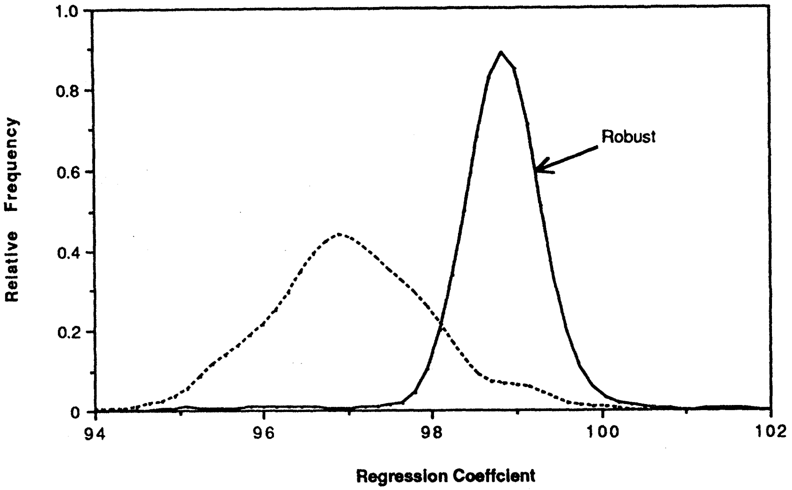

Figure 6: Bootstrap distributions of estimated linear coefficients in OLS and robust regression fits.

simulations of Efron (1982: Table 5.2), standard errors from the delta method are too small, particularly for statistics such as the correlation.

Applying the delta method to indirect effects is not as easy as in the scalar case described above because $\hat{\theta}$ now consists of a vector of regression coefficients. As a result, the linear approximations (8), (9), and (10) require vector calculus. For example, if $\hat{\theta}$ denotes the vector of coefficient estimates from our model, then $g(\hat{\theta})$ is typically a product such as $\hat{\theta}_2$, $\hat{\theta}_4$, $\hat{\theta}_5$, and the scalar derivative becomes a vector of partial derivatives; details appear in Sobel (1982).

The bootstrap can also be applied to indirect effects. One simply forms the coefficient estimates from the various equations of the model and computes the indirect effect. Each equation in the model is bootstrapped B times using the *same* B bootstrap samples for all of the equations. In the simple case of a recursive model (i.e., one with no feedback), the collection of bootstrap indirect effects are just the products of the bootstrapped regression coefficients from the several equations. Because the bootstrap regression estimates are based on the same B bootstrap samples, the coefficient estimates are correlated across the equations, as they should be.

In general, the delta method and bootstrap give similar results for indirect effects. Bootstrap standard errors are generally larger than

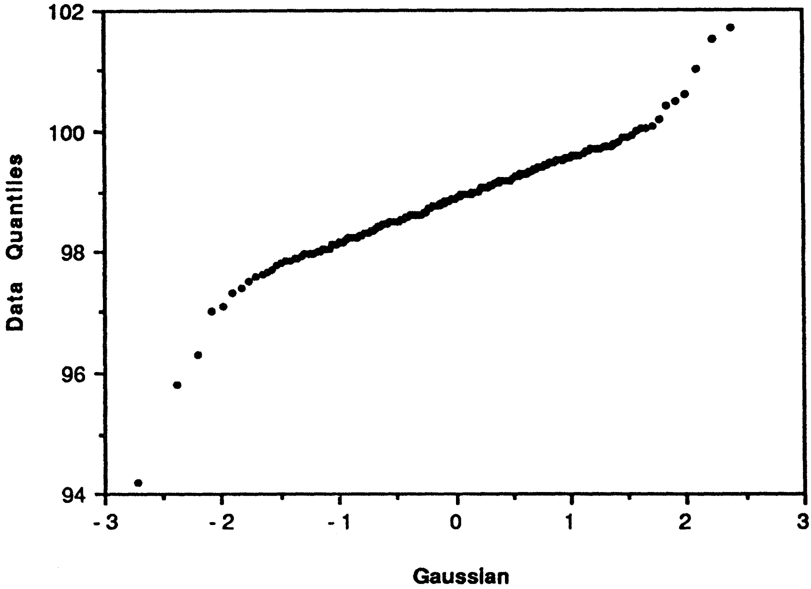

Figure 7: Quantile comparison plot of the bootstrap replications of the robust slope estimator reveals heavy tails.

those from the delta method, with the size of the difference depending on the sample size. More importantly, however, the bootstrap can reveal skewness in the distribution of an estimated indirect effect. Because the delta method is teamed with a Gaussian distribution, it does not reveal asymmetries. The distributions in Figure 8 are the smoothed bootstrap and delta-method approximations to the distribution of an indirect effect in a small, recursive path model estimated from a sample of 50 observations (Bollen and Stine, 1988). The two distributions are similar, but the bootstrap suggests asymmetry that the normal approximation associated with the delta method cannot capture. In the same model with a larger sample size (n=172), the differences are quite small.

BOOTSTRAP CONFIDENCE INTERVALS

A substantial body of recent research in statistics concerns bootstrap confidence intervals. Rather than bury the reader in the details of the

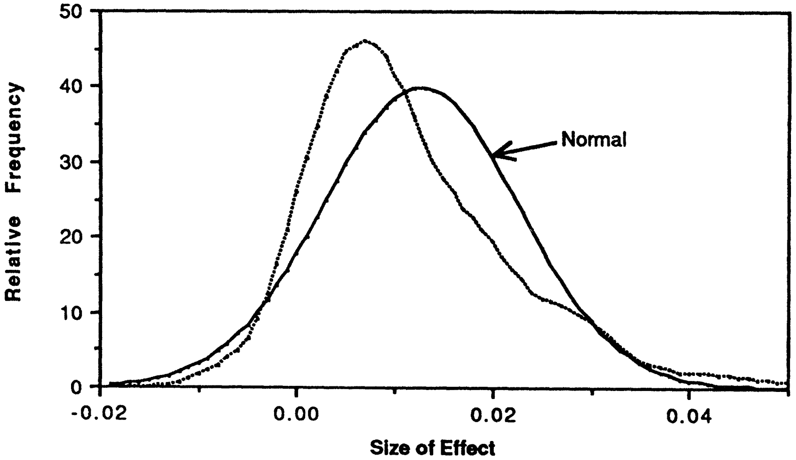

Figure 8: Bootstrap and normal approximations to the distribution of an indirect effect in a small path model.

most recent advances, this section displays the key ideas that underlie this research. This section begins with a quick overview of bootstrap t-intervals, which are a variation of intervals based on the classic t-statistic. Treatment of percentile intervals follows. Percentile intervals are closely related to the smoothed histograms of the bootstrap replications shown in the regression examples.

BOOTSTRAP t-INTERVALS

Bootstrap t-intervals share the form of the classic t-interval, but do not require the Gaussian populations or the use of a t-table for critical values. Essentially, a new table constructed using bootstrap replications replaces the familiar t-table in each application. The case of a confidence interval for the mean of a Gaussian population illustrates the ideas.

The usual 90% confidence interval for the mean μ of a normal population based on a sample of size n is $[\bar{x} \pm t(.05, n-1)\ s/\sqrt{n}]$, where $t(\alpha, df)$ is the α percentile of Student's t-distribution with df degrees of freedom and s is the sample standard deviation. The validity of this interval (i.e., the reason that it really is a 90% confidence interval for μ) relies upon the fact that

$$\Pr\{\sqrt{n}\,(\bar{x}-\mu)/s \leq t(\alpha,df)\} = \alpha, \quad 0<\alpha<1 \qquad [11]$$

The usual interpretation of the probability in (11) implies that the ratio $\sqrt{n}\,(\bar{x}-\mu)/s$ is less than the critical value $t(\alpha,df)$ in 100α% of the utopian collection of samples from the Gaussian population. The bootstrap t-interval is based in the same logic, and seeks an analogous value $t^*(\alpha;\, n)$ such that[16]

$$\Pr^*\{\sqrt{n}\,(\bar{x}^*-\bar{x})/sd_B{}^*(\bar{x}^*) \leq t^*(\alpha;n)\} = \alpha, \quad 0<\alpha<1 \qquad [12]$$

In (12) probabilities associated with bootstrap sampling from the data replace probabilities associated with sampling from the true population. This change replaces "Pr" in (11) with "Pr*," which denotes the probability induced by bootstrap resampling from the empirical distribution F_n. Also, $sd_B{}^*(\bar{x}^*)$ is the bootstrap estimate of the standard deviation of the sample mean based on B replications $\bar{x}^{*(1)}, \ldots, \bar{x}^{*(B)}$. Rather than find a percentile from the t-table, the bootstrap approach is to find it directly from the distribution of the ratio $(\bar{x}^*-\bar{x})/sd_B{}^*(\bar{x}^*)$. For example, the 90% bootstrap t-interval for μ is

$$[\bar{x}+t^*(.05;\, n)\times sd_B{}^*(\bar{x}^*),\ \bar{x}+t^*(.95;\, n)\times sd_B{}^*(\bar{x}^*)]$$

Finding $t^*(\alpha;n)$ requires simulation, as the mathematics quickly become intractable. Hall (1986a) gives an example of the mathematics involved.

To find $t^*(\alpha;n)$ requires a nested bootstrap simulation. Each iteration of the outer loop of the simulation generates a bootstrap replication of the pivot $R^* = (\bar{x}^*-\bar{x})/sd_B{}^*(\bar{x}^*)$. The required value of $t^*(\alpha;n)$ is the α percentile of the simulated collection of pivots. The inner loop is needed to find the bootstrap standard error estimate. The algorithm is:

(1) Draw B_1 bootstrap samples from the original observations and denote these by $X^{*(j)} = (x_1{}^{*(j)}, \ldots, x_n{}^{*(j)}),\ j = 1, \ldots, B_1$.

(2) For each of these B_1 bootstrap samples, estimate the standard deviation of the mean by bootstrapping:

(2a) Draw B_2 bootstrap samples from $X^{*(j)}$ and label these

$$X^{*(jb)} = (x_1{}^{*(jb)}, \ldots, x_n{}^{*(jb)}),\ b = 1, \ldots, B_2$$

where $x_i{}^{*(jb)}$ is sampled with replacement from the observations in the sample $X^{*(j)}$.

(2b) Estimate the mean $\bar{x}^{*(jb)}$ of each of the B_2 bootstrap samples $X^{*(jb)}$

(2c) Compute $sd_{B2}{}^{*(j)}$ from the collection of bootstrap means

$$sd_{B2}{}^{*(j)}(\bar{x}^*) = [\sum_{b=1}^{B_2} (\bar{x}^{*(jb)} - \bar{x}^{*(j\cdot)})^2 / (B_2 - 1)]$$

where $\bar{x}^{*(j\cdot)}$ is the average of the B_2 bootstrap means.

(3) Form the bootstrap pivot $R^{*(j)} = (\bar{x}^{*(j)} - \bar{x}) / sd_{B2}{}^{*(j)} (\bar{x}^*)$ for each of the B_1 initial bootstrap samples.

(4) Use the collection of bootstrap pivots to obtain the desired percentiles

$$t^*(\alpha;n) = \alpha \text{ quantile of the } R^{*(j)}$$

At the cost of much more calculation, one gains the freedom of not having to know the distribution of $\sqrt{n}\,(\bar{x}-\mu)/s$.[17]

Fortunately, only a relatively small simulation is needed to obtain the optimal level of accuracy. If the number of bootstrap samples B_1 in the outer loop of this nested simulation is roughly equal to the sample size, then the difference in coverage from performing an infinite amount of resampling is very slight. Such accuracy requires careful choice of the level α. Because B_1 replications of the ratio R^* divide the line into B_1+1 segments, α and B_1 should be chosen so that $\alpha = k/(B_1+1)$ for some positive integer k. For example, to get a 90% interval, the smallest satisfactory number of replications B_1 is 19 because these divide the line into 20 segments, each holding 5% of the probability. If $R_{(1)}{}^* \le R_{(2)}{}^* \le \ldots R_{(19)}{}^*$ are the ordered $R^{*(j)}$'s, then the lower endpoint of the bootstrap interval is $\bar{x} + R_{(1)}{}^* sd_B{}^*(\bar{x}^*)$ and the upper is $\bar{x} + R_{(19)}{}^* sd_B{}^*(\bar{x}^*)$. The cost of doing so little resampling lies in the length of the interval. Although the coverage accuracy of the interval is hardly affected by B_1, the length of the interval is. Doing too little resampling generally leads to intervals that are unnecessarily long (Hall, 1986b).

CONFIDENCE INTERVALS FOR THE CORRELATION

Building a confidence interval for the correlation of two random variables is neither so simple nor obvious as doing so for the mean. Many of the problems are suggested by the bootstrap distribution of the correlation of the law school data shown in Figure 9. These data appear in many of the Efron references (e.g., Efron, 1982) and are grade-point averages (GPA) and law school aptitude test (LSAT) scores from 15 U.S. law schools. The sample correlation is rather large, $r = 0.776$. The bootstrap distribution of the correlation (B=1000) shown in Figure 10

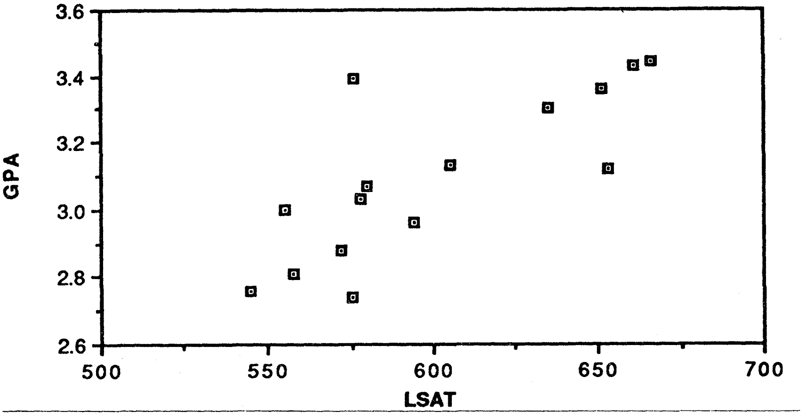

Figure 9: Efron's law school data.

is skewed. Consequently, symmetric t-intervals are inaccurate. For example, an approximation to the standard error of r is $SE(r) \approx (1-r^2)/(n-2)^{1/2}$, so that an approximate 90% interval for the population correlation ρ is

$$[0.776 + 1.645 \times 0.115] = [0.776 + 0.189] = [0.587, 0.965].$$

The interval gives no indication of the skewness revealed in the bootstrap distribution and relies upon an unrealistic Gaussian approximation. In fact, the upper endpoint of the 95% interval (replace 1.645 with 1.96) is greater than 1. The approximation to SE(r) also reveals a further complication: The variability of r depends on the value of ρ. The larger ρ becomes, the less variable r is. Getting a better interval depends, in part, on how well one can transform the correlation into a statistic that has a normal distribution whose variance does not depend on the underlying parameter.

The required transformation in this case is Fisher's z-transformation, $\phi(r) = (1/2) \ln\{(1+r)/(1-r)\}$. This transformation maps the range of the correlation $[-1, +1]$ onto the whole line, $-\infty < \phi(r) < +\infty$. In so doing, it removes much of the asymmetry seen in Figure 10. Also, the variance of $\phi(r)$ is approximately $1/(n-3)$, which does not depend upon ρ. Because the distribution of $\phi(r)$ is more nearly a Gaussian distribution, we can construct a t-interval for $\phi(\rho)$. We can then use this interval to get one for ρ. The idea, then, is to form an interval on a transformed scale where the usual strategy of [estimate±t×std. err.] is roughly correct.

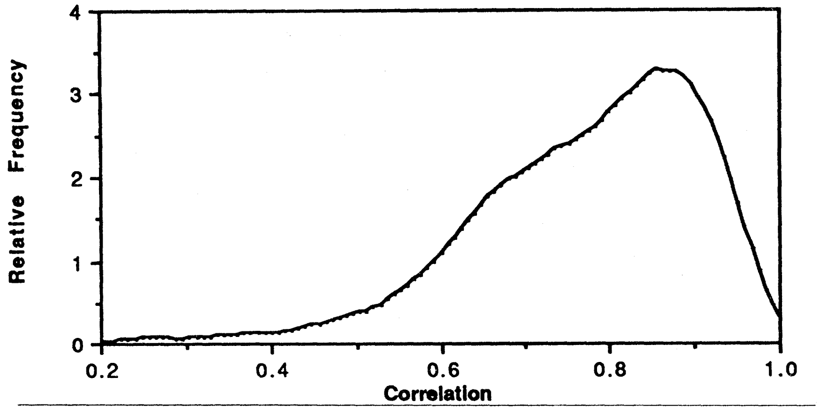

Figure 10: Bootstrap distribution of the correlation of LSAT and GPA.

Then we reverse, or *invert*, the transformation to get back to the original scale and finish with an asymmetric interval.

An example using the law school data illustrates these ideas. Using Fisher's transformation gives $\phi(0.776) = (1/2)\ln(1.776/0.224) = 1.035$, and an approximate 90% confidence interval for $\phi(\rho)$ is

$$[1.035 \pm 1.645 \times 0.289] = [0.560, 1.510]$$

To get an interval for ρ, note that if $\phi(r) = z$, then solving for r in terms of z gives $r = (e^z - e^{-z})/(e^z + e^{-z}) = \tanh(z)$, a function found on some hand calculators. Applying this transformation to the endpoints of the interval for $\phi(\rho)$ yields the desired interval for ρ

$$[0.508, 0.907] = [0.776-0.268, 0.776+0.131].$$

This interval is asymmetric and cannot include values outside the range -1 to $+1$.

Bootstrap percentile intervals automatically accomplish much of what Fisher's z-transformation does. Intervals for the correlation based on Fisher's z transformation work rather well even for fairly large values of ρ, but their use requires that we know about this transformation. Generally, it is rather hard to find a transformation for an arbitrary statistic that works as well as ϕ works for the sample correlation, although some recent work seeking to automate this search shows promise (Tibshirani, 1988). Bootstrap percentile intervals are more direct. Suppose that we computed B bootstrap replications of the sample correlation $r^{*(1)}, r^{*(2)}, \ldots, r^{*(B)}$. In this case, B will need to be rather large—on the order of 1000. As with random resampling in regression, we resample from the pairs ($LSAT_i$, GPA_i) so as to preserve the relationship between the two variables. The 90% bootstrap percentile interval is then $[r^*(0.05), r^*(0.95)]$, where $r^*(p)$ $(0<p<1)$ is the 100*p*th percentile of the bootstrap distribution of the correlation. That is, we merely sort the bootstrap replicates, find the one greater than 5% of the r^*'s, and use it for the lower endpoint. Similarly, the replicate that is greater than 95% of the r^*'s becomes the upper endpoint.

The key property of percentile intervals is that, in a sense, they automatically make use of Fisher's transformation. Suppose that we knew of the skewness in the distribution of the sample correlation and used Fisher's z-transformation *with* the bootstrap. Instead of looking at 1000 replications of r, we instead would compute 1000 replications of $\phi(r)$. The resulting bootstrap percentile interval for $\phi(\rho)$ would then be formed from the 5th and 95th percentiles of the collection of $\phi(r^*)$'s. When this interval is inverted to give an interval for ρ, however, we get the same interval that we would have gotten without the transformation. This invariance occurs because ϕ is a monotone function (it steadily increases). Thus, it preserves the order of the bootstrap replications. The 5th percentile of the transformed values $\phi(r^*)$ is simply ϕ applied to the 5th percentile of the r^*'s. When the percentile interval $[\phi\{r^*(0.05)\}, \phi\{r^*(0.95)\}]$ is inverted back to the original scale, we obtain the same interval as before, $[r^*(0.05), r^*(0.95)]$. How well this automatic transformation performs depends on some key assumptions laid out in the next section.

PERCENTILE AND BIAS CORRECTED PERCENTILE INTERVALS

The validity of the percentile interval stems from the existence of a normalizing transformation. Suppose that we want a confidence interval for some population parameter θ that has been estimated by the statistic $\hat{\theta}$. The simplest, but most restrictive assumption that guarantees that percentile intervals give the correct coverage is to assume

$$\begin{matrix} (\hat{\theta}^* - \hat{\theta}) \\ \quad \sim N(0, \sigma^2) \\ (\hat{\theta} - \theta) \end{matrix} \qquad [A1]$$

That is, the bootstrap analogy holds and both deviations from "true" values have a normal distribution with mean 0. Of course, the bootstrap distribution of $\hat{\theta}^*$ is discrete and cannot be normal, so we interpret (A1) to mean that the bootstrap distribution becomes very close to normal as the sample size increases.

To see why (A1) implies that percentile intervals give the correct coverage, we have to compare the usual Gaussian interval for θ with the bootstrap percentile interval and show that they are the same. Begin by noting that if σ is known, then a 1-2α confidence interval for θ is $[\hat{\theta}+z(\alpha)\sigma,\ \hat{\theta}+z(1-\alpha)\sigma]$. Again, $z(\alpha)$ is the α percentile of the Gaussian distribution; for example, $z(.05) = -1.645$ and $z(.95) = 1.645$. Under bootstrap resampling, the probability of a bootstrap replicate $\hat{\theta}^*$ being less than the lower endpoint of this interval is

$$\begin{aligned} G^*\{\hat{\theta}-z(\alpha)\sigma\} &= \Pr{}^*\{\hat{\theta}^* \le \hat{\theta}-z(\alpha)\sigma\} \\ &= \Pr{}^*\{(\hat{\theta}^*-\hat{\theta})/\sigma \le z(\alpha)\} \\ &= \alpha \end{aligned}$$

because (A1) implies that $(\hat{\theta}^*-\hat{\theta})/\sigma$ has a standard normal distribution. Thus, the value that cuts off the lower $100\alpha\%$ of the bootstrap distribution of $\hat{\theta}^*$, the lower endpoint of the bootstrap interval, is also the lower endpoint of the standard interval. As a result, the bootstrap interval $[G^{*-1}(\alpha),\ G^{*-1}(1-\alpha)]$ has the correct coverage.[18]

Assumption (A1) is rather restrictive, and the first generalization allows for a transformation to normality. The statistic $\hat{\theta}$ is allowed to have a non-Gaussian distribution, and it is assumed that an invertible transformation to normality exists, like Fisher's transformation of the correlation. If this normalizing transformation is labelled "h," then the generalization of (A1) is to assume

$$\begin{array}{l} (h(\hat{\theta}^*)-h(\hat{\theta})) \\ \qquad\qquad \sim N(0,\sigma^2) \\ (h(\hat{\theta})-h(\theta)) \end{array} \qquad [A2]$$

Under (A2), the usual $1-2\alpha$ interval for $h(\theta)$ is

$$[h(\hat{\theta})+z(\alpha)\sigma,\ h(\hat{\theta})+z(1-\alpha)\sigma].$$

Using the existence of the inverse transformation h^{-1}(tanh in the case of Fisher's z-transformation), the interval for θ is $[h^{-1}\{h(\hat{\theta})+z(\alpha)\sigma\},$

$h^{-1}\{h(\hat{\theta})+z(1-\alpha)\sigma\}$]. To arrive at this interval requires knowing h, h^{-1}, and σ. Again, the percentile interval arrives at the same endpoints without requiring so much *a priori* information. Arguing as before, the probability of $\hat{\theta}^*$ being less than the lower endpoint of the desired interval is

$$\begin{aligned} \Pr{}^*\left\{\hat{\theta}^* \leq h^{-1}(h(\hat{\theta})+z(\alpha)\sigma)\right\} &= \Pr{}^*\left\{(h(\hat{\theta}^*)-h(\hat{\theta}))/\sigma, \leq z(\alpha)\right\} \\ &= \alpha \end{aligned}$$

and the percentile interval gets the correct coverage *without* having to be told the normalizing transformation.

As broad as assumption (A2) first appears, it does not obtain in some common situations, particularly in the presence of bias. To get a sense for why the percentile interval fails, consider what happens if it is used with a biased estimator. Suppose that the estimator $\hat{\theta}$ is biased for the true parameter θ and tends to be too small—say $\hat{\theta}$ is on average 2 less than θ, $E(\hat{\theta}-\theta) = -2$. If the bootstrap analogy holds, then the bootstrap replicates $\hat{\theta}^*$ are also biased for the true parameter of the bootstrap population, $\hat{\theta}$. Thus, the bootstrap replicates used to form the endpoints of the percentile interval are shifted to the left of $\hat{\theta}$, when in fact they should be shifted to the right toward θ.

Fortunately, a diagnostic method exists that measures discrepancies from assumption (A2) and provides the means to correct the problem. This diagnostic is to check that the probability of $\hat{\theta}^*$ being less than $\hat{\theta}$ is 1/2; that is, half of the bootstrap values $\hat{\theta}^*$ should be less than the observed statistic $\hat{\theta}$, a condition known as median unbiased. Under (A2), half of the $\hat{\theta}^*$'s should be less than $\hat{\theta}$ because it is assumed that $h(\hat{\theta}^*)$ is normally distributed about $h(\hat{\theta})$. Using the notation of bootstrap distributions, we need to check that $G^*(\hat{\theta}) = \Pr^*(\hat{\theta}^* \leq \hat{\theta}) = 0.5$. For the law school data, only 446 of the 1000 bootstrap replications of the correlation are less than the observed correlation, $G_{1000}{}^*(r) = 0.446$. Because $G_{1000}{}^*$ is an estimate of G^*, we need to decide whether the observed deviation from 0.5 is indicative of a problem or merely the result of sampling fluctuations. In other words, we need to test the null hypothesis H_0: $G^*(r) = 0.5$. Under H_0, the number of the 1000 bootstrap replicates r^* that are less than r has a binomial distribution with mean 500 and standard deviation $0.5 \times 1000^{1/2} = 15.8$. Because the observed count is well over 3 standard deviations from 500, assumption (A2) does not hold.

Bias-corrected percentile intervals allow for the presence of bias and consequently lead to a more general bootstrap confidence interval.

Rather than require that $h(\hat{\theta}^*) - h(\hat{\theta})$ and $h(\hat{\theta}) - h(\theta)$ both be centered on 0, these intervals allow for some bias. The assumed behavior is

$$\begin{array}{l} (h(\hat{\theta}*)-h(\hat{\theta})) \\ \qquad\qquad\qquad \sim N(-Z_0\sigma, \sigma^2) \qquad\qquad [A3] \\ (h(\hat{\theta})-h(\theta)) \end{array}$$

where the bias is expressed as a constant, $-Z_0$, number of multiples of the standard deviation σ. Because (A3) implies that the mean of $h(\hat{\theta})$ is $h(\theta) - Z_0\sigma$, it follows that

$$\{h(\hat{\theta})-h(\theta) + Z_0\sigma\}/\sigma \sim N(0,1)$$

and the standard $(1-2\alpha)$ interval for θ is

$$[h^{-1}\{h(\hat{\theta})+Z_0\sigma+z(\alpha)\sigma\},\ h^{-1}\{h(\hat{\theta})+Z_0\sigma+z(1-\alpha)\sigma\}]$$

This interval requires that we know the transformation h, its inverse h^{-1}, *and* both Z_0 and σ. But again, if we consider the probability of $\hat{\theta}^*$ being less than the lower endpoint of the desired interval, we find that the lower endpoint of the normal-theory interval is related to a percentile of the bootstrap distribution

$$\begin{aligned} G^*\{h^{-1}(h(\hat{\theta})+Z_0\sigma+z(\alpha)\sigma)\} &= Pr^*\{h(\hat{\theta}^*)\le h(\hat{\theta})+Z_0\sigma+z(\alpha)\sigma\} \\ &= Pr^*\{(h(\hat{\theta})^*-h(\hat{\theta})+Z_0\sigma)/\sigma\le 2Z_0+z(\alpha)\} \\ &= \Phi(2Z_0+z(\alpha)) \end{aligned}$$

where $\Phi(x)$ is the cumulative normal distribution, $\Phi(x) = Pr\{N(0,1)\le x\}$. The presence of the bias implies that the percentile interval constructed using the α and $1-\alpha$ percentiles is no longer correct. Instead the lower endpoint of the bias-corrected bootstrap interval needs to be the $\Phi(2Z_0+z(\alpha))$ percentile, which suggests that we still need to know the bias factor Z_0. However, the same diagnostic that suggests the need for a bias adjustment gives an estimate of Z_0. The proportion of bootstrap replicates $\hat{\theta}^*$ that are less than $\hat{\theta}$ is the proportion of the normal distribution less than Z_0

$$G^*(\hat{\theta}) = Pr^*\ \{(h(\hat{\theta})-h(\theta)+Z_0\sigma)/\sigma\le Z_0\} = \Phi(Z_0)$$

Thus, $Z_0 = \Phi^{-1}\{G^*(\hat{\theta})\}$, the z value corresponding to $Pr^*\{\hat{\theta}^*\le\hat{\theta}\}$. Notice that the bias corrected interval reduces to the usual percentile interval if $\hat{\theta}^*\le\hat{\theta}$ in half of the samples. In this case, $Z_0 = 0$ because $\Phi(0.5) = 0$.

Bias correction makes a slight difference in the interval for the correlation. The effect of bias correction in this example is subtle because

$$\begin{aligned} Z_0 &= \Phi^{-1}\text{ (proportion of BS correlations less than r)} \\ &= \Phi^{-1}(0.446) = -0.13 \end{aligned}$$

For the lower endpoint, $\Phi\{2Z_0+z(0.05)\} = \Phi\{2(-0.13)-1.65\} = 0.028$. Thus, the value cutting off the lower 2.8% of the bootstrap distribution of r^* becomes the lower endpoint of the 90% interval, rather than the 5% point. For the upper endpoint, the 91.8% point of G^* is used. Although the interval endpoints appear to imply that the coverage is no longer 0.9 (because 0.918–0.028=0.89), the interval is nonetheless an estimated 90% confidence interval. Table 6 summarizes the several types of intervals. Fisher's transformation produces an asymmetric interval; the direct normal approximation does not. The bootstrap intervals become progressively more skewed as one moves down the table.

ACCELERATED BOOTSTRAP INTERVALS

Efron (1987) proposes *accelerated percentile intervals* as further enhancement of percentile intervals. As illustrated in Table 6, accelerated intervals can be much more asymmetric than the bias-corrected interval. For some familiar statistics, including the sample variance s^2, the transformation needed for the bias-corrected interval does not exist because the variance of the normal approximation depends on the value of θ (Schenker, 1985; Efron, 1987). Simulations of an interval for the variance σ^2 based on samples of 35 Gaussian observations (Schenker, 1985) revealed that the coverage of the 90% percentile interval was much too small, only 82%, and the coverage of the bias-corrected percentile interval was not much better—only 85%.

Like the bias-corrected interval, accelerated bootstrap intervals alter the percentiles of the bootstrap distribution that are used for the endpoints of the bootstrap interval. For the correlation in the law school data, the bias-corrected percentile interval is too far to the right. Accelerated intervals remedy much of this problem by using as endpoints the 1st and 89th percentiles of the bootstrap distribution of the correlation, as compared to the 3rd, and 92nd used in the bias-corrected interval. The accelerated interval thus reaches farther into the tail of the distri-

TABLE 6
Several 90% bootstrap confidence intervals for the correlation of the law school data of 15 observations with sample correlation 0.776.

Method	Interval
Classical	
Without Fisher's z	[0.56, 0.99]
With Fisher's z	[0.49, 0.90]
Bootstrap (B=1000)	
Percentile	[0.55, 0.94]
BC percentile	[0.52, 0.93]
Accelerated percentile	[0.43, 0.92]

bution. Unfortunately, the computation of the accelerated intervals is more involved than that of the percentile intervals and will not be covered further here. Details of the calculations appear in Efron (1987) and DiCicco and Tibshirani (1987).

BOOTSTRAP PREDICTION INTERVALS

A variation on utopian sampling suggests how to use the bootstrap to find confidence intervals for predictions. Having estimated the regression model (1), we frequently forecast the values of new observations $y_f = x_f'\beta + \varepsilon_f$ with the predictor $x_f'\hat{\beta}$. To construct an interval that measures the uncertainty of this forecast, the standard approach is to assume that the errors in the regression model possess a Gaussian distribution. If k regressors are used in the model, then a prediction interval with coverage $1-2\alpha$ for y_f is $I_G(f) = [x_f'\hat{\beta}+t(\alpha;\, n-k)s_f,\ x_f'\hat{\beta}+t(1-\alpha;\, n-k)s_f]$ where s_f is the standard error of the forecast, $s_f^2 = s^2\,(1+x_f'\,(X'X)^{-1}x_f)$.[19]

The idea behind bootstrap prediction intervals is to replicate the entire sampling process and directly observe the prediction error. This procedure consists of generating bootstrap replicates of the observations Y^* and X^*, using the fitted model to obtain a future value $y_f^* = x_f'\hat{\beta} + e_f^*$, and measuring the observable prediction error $PE_f^* = y_f^* - x_f'\hat{\beta}^*$. Here, e_f^* is a random draw from the empirical distribution of the residuals,

TABLE 7
Coverage of bootstrap and normal theory prediction intervals for regression models with errors from several distributions.

Error Distribution	Bootstrap	Normal Theory
Gaussian	0.78	0.80
Logistic	0.79	0.81
Student's t (4 df)	0.78	0.83

which is independent of the resampling used to generate $\mathbf{Y}^*$ and $\mathbf{X}^*$. The resulting interval resembles a percentile interval; the percentiles of the bootstrap prediction errors are added to the original prediction rather than a parameter estimate. If we let $H^{*-1}(\alpha)$ denote the 100α percentile of $PE_f^{*(1)}, \ldots, PE_f^{*(B)}$, then a $1-2\alpha$ coverage bootstrap prediction error for y_f is $I_{BS}(f) = \{x_f'\hat{\beta} + H^{*-1}(\alpha), x_f'\hat{\beta} + H^{*-1}(1-\alpha)]$.

Table 7 contrasts the coverage probabilities of the bootstrap prediction intervals with the usual normal theory intervals. Although the coverage of I_{BS} is slightly less than 0.9, the coverage is consistent for all three distributions. One can prove in special cases that the distribution of the coverage of I_{BS} is asymptotically distribution-free: The coverage of the bootstrap intervals does not depend on the shape of the underlying population. In contrast, the coverage of the normal theory interval deviates further from the nominal amount as the distribution becomes more long-tailed. Further details and a computational enhancement appear in Stine (1985); similar methods for time-series models appear in Stine (1987).

COMPUTING BOOTSTRAP ESTIMATES

It would be convenient at this point to be able to direct the reader to a well-developed commercial software package that included bootstrap procedures. But because such a package does not exist, it is useful to remember some computational issues that arise in resampling. The bootstrap calculations that appear in this article were prepared using a collection of APL programs written by the author. It is also possible to use the SAS macrolanguage to write special routines to do bootstrap calculations. One can write such a bootstrapping macro in any statistical package that permits the user to make function calls to statistical routines and supports random number generation.

SOME GENERAL POINTS TO REMEMBER

Most interesting applications of the bootstrap, such as confidence intervals, require simulation. Although no rules exist to always give the best computing strategy, a few general points deserve emphasis. Some have already been mentioned, but are repeated here.

(1) *Simulation is not always necessary.* The bootstrap is not a collection of simulation algorithms. Rather, the bootstrap is a methodology based on substituting the sample for the unknown population. Often, simple mathematics can replace simulation, as in the example of computing the variance of the sample average.

(2) *Improve naive bootstrap methods with substantive knowledge.* The naive bootstrap prediction intervals of the earlier section are intuitive, but the resulting computational strategy is not computationally efficient. A basic understanding of the structure of predictions leads to an algorithm that produces more accurate intervals with less resampling (Stine, 1985).

(3) *Avoid iterating nonlinear statistics.* Many nonlinear estimators, including robust regression, begin at some starting point and sequentially improve the solution. Although these are often iterated "until convergence," one step of such a method is generally sufficient because one step from a consistent initial estimate is asymptotically efficient (Zacks, 1971, sec. 5.5). Related ideas appear in Jorgensen (1987).

(4) *Use the same bootstrap samples when comparing estimators.* To make the most of bootstrap comparisons, use the same bootstrap samples for both estimators. If different bootstrap samples are used for comparing two estimators, some of the differences between the estimators will be due to differences in the bootstrap samples. Using the same samples induces correlation, which helps comparisons.

HOW MANY BOOTSTRAP SAMPLES ARE NECESSARY

One of the most common questions about using the bootstrap is, How many bootstrap samples are needed? The answer depends upon the problem, but B on the order of 100 is typically needed for standard error estimates, whereas B $\approx$ 1000 or larger is typically necessary for estimating a percentile of a distribution. Even with an infinite number of bootstrap replications, the bootstrap standard error is still a random variable. If B is chosen by these rough guidelines, sample-to-sample variation in the bootstrap standard error is typically much larger than the variation induced by limiting the size of the simulation. Tibshirani (1985) gives a precise description of how to determine B when estimating standard errors and percentiles, and further ideas appear in Efron (1987, sec. 9).

FUTURE DIRECTIONS

Several recent applications of bootstrap methods reach beyond straightforward applications in variance estimation and confidence intervals. One can expect to see further extensions along these lines. Finally, a closing warning shows that the bootstrap need not give the correct answer, especially when the model imposed on the data is incorrect.

NONPARAMETRIC REGRESSION METHODS

Inexpensive computing resources have renewed interest in nonparametric regression. In nonparametric regression, the conditional expectation of the response Y is not restricted to the linear form of model (1), but is instead permitted to be an arbitrary function of X. Several methods exist for fitting such functions, and one is based on smoothing splines. The usual spline function is a smooth curve that interpolates the data; a smoothing spline is a related function, but it does not pass through every observation. As an example, Figure 11 shows a smoothing spline fit to the age and skeletal age of 100 black male adolescents in a study of hypertension in blacks (Katz et al., 1980). Skeletal age is a measure of maturation based on interpreting x-rays of the hand and wrist. What is interesting in the figure is the location of the bend in the curve between 13 and 14 years of age. Does the location of the kink really fall in this interval or are we being misled by sampling variation?

The bootstrap offers one approach to answering this question. Simply generate bootstrap samples from the 100 observations, fit smoothing splines to each, and see how the location of the kink varies. Figure 12 shows smoothing splines fit to five bootstrap samples. Although considerable variation exists in the fitted splines before age 14, the curves come together at about age 14 and all flatten out. Maturation, as measured by skeletal age, appears to stop in this sample after age 14.

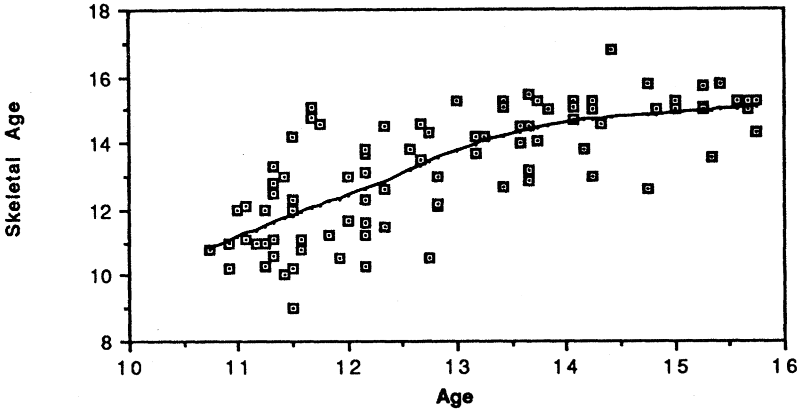

Figure 11: Smoothing spline fit to skeletal age growth data.

Further details on smoothing splines appear in Silverman (1985), and other examples of nonparametric regression appear in Gasser et al. (1984).

MODEL ASSESSMENT AND ERROR RATES

The bootstrap is a useful tool for evaluating overall model fit. Summaries such as s^2 and R^2 use deviations of the data from the fitted model to measure the success of the model. Such criteria are often "optimistic" because the same data that generated the model are used to assess the model. Models based on one sample often fit new data poorly. Adjustments for degrees of freedom offer some improvement, but the bootstrap suggests how to go further. Efron (1983) uses the bootstrap to estimate how optimistic the standard estimates of goodness-of-fit are.

The basic idea is simple: Evaluate models constructed from bootstrap samples based on how well they fit the bootstrap population, the original sample. One begins by constructing a predictive model from a bootstrap sample. The model could be logistic regression, discriminant analysis, or even linear regression. These models provide an estimate of how well they can predict future observations from the same population, such as the classification error rate in discriminant analysis. Because the bootstrap population is known, we can see how well the estimated model predicts the population. By comparing the actual accuracy to the model's claimed accuracy, we get an idea of how well

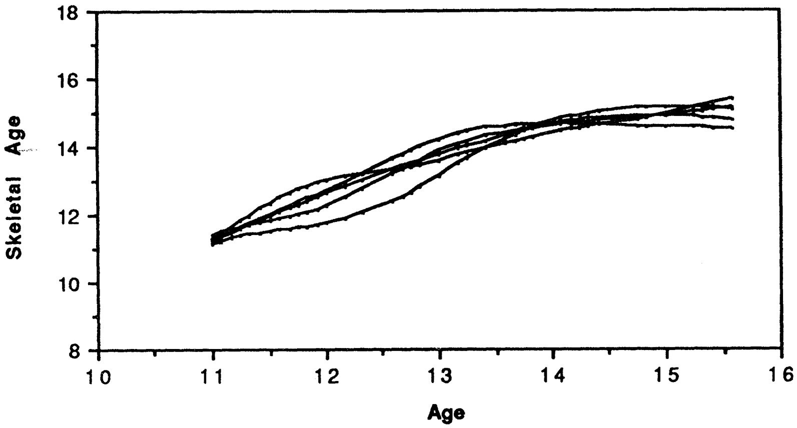

Figure 12: Replications of the smoothing spline from five bootstrap samples of the growth data

the goodness-of-fit measure performs. Once it is known that the model is never so accurate as it claims to be, we can inflate the size of the expected error by an amount suggested by the bootstrap.

CONFIDENCE INTERVALS

Confidence intervals in very small samples remain a problem. When the data supply little information, the data analyst has to impose some form of external structure. Rather than appeal to the existential transformation of bootstrap percentile intervals, handle skewness by finding a normalizing (or at least symmetrizing) transformation. The bootstrap can help in this search: Use the bootstrap distribution of the statistic to judge the effectiveness of a given transformation. The methods of exploratory data analysis are also useful in this search, noting that now one is seeking to symmetrize the distribution of a statistic rather than the distribution of the sample.

The confidence intervals described in the section on bootstrap confidence intervals apply to scalar-valued statistics. What if we want to form a simultaneous confidence region? For example, it is common in regression to form a confidence region for several of the slope parameters simultaneously. Bootstrap intervals are not so well established in this area. Once we look simultaneously at several parameters, it becomes hard to generate enough bootstrap replications to get reasonable percentile intervals, a problem that is similar to density estimation over the plane (Silverman, 1986). Also, it is not clear just how we should define a bootstrap confidence region. Some recent work on this problem appears in Hall (1987).

HYBRID ESTIMATORS

The bootstrap comparison of the robust and least-squares estimators suggests constructing a new hybrid estimator. A hybrid estimator is a mixture of several estimators, such as least-squares and robust estimators. When the bootstrap standard errors reveal that the OLS estimator is more stable, then the hybrid is the least-squares estimator. If the robust estimator seems more stable, then it is to be the value of the hybrid. An early example of this idea is Switzer's adaptive trimmed mean. This estimate of location selects the amount of trimming by minimizing the *jackknife* estimate of variance. (See Efron (1982: 28) for further discussion of this estimator and comparisons to other techniques.)

OTHER TOPICS

The bootstrap has found applications in virtually every area of statistics. For example, censored data occur when we do not observe the actual value of some variable, but only know that it exceeds some known cutoff. Such data problems appear in event-history analysis. Efron (1981) showed how to use the bootstrap to estimate sampling properties of the Kaplan-Meier estimator, which appears in the analysis of censored data, and further applications appear in Akritas (1986). Applications in time-series analysis are less common, primarily because serial correlation requires assumptions about the structure of the data. Freedman (1984) describes bootstrapping in very complex econometric models and includes proofs of the large sample validity of the bootstrap in this setting. Examples of bootstrapping with time-series models appear in Efron and Tibshirani (1986), Swanepoel and Van Wyck (1985), and Stine (1987). The bootstrap is also useful in multivariate analysis, with applications ranging from the variation of principal component weights (Diaconis and Efron, 1983) to error rates in discriminant analysis (Efron, 1983).

All of our applications have treated the data as if they are a simple random sample. Real data are seldom so simple, and are often gathered through complex sampling designs. In these cases, naive bootstrap methods do not replicate the actual sampling structure of the data, and they give incorrect results. Rao and Wu (1988) address these issues for a variety of sampling designs, including stratified cluster sampling and two-stage cluster sampling.

PROBLEMS IN PARADISE: A SITUATION IN WHICH THE BOOTSTRAP FAILS

The bootstrap does not always yield the correct standard error estimator, particularly if the resampling scheme does not parallel the structure of the actual sampling mechanism. The presence of correlated observations presents a case in which it is easy to misuse the bootstrap and obtain misleading results. Once the assumption of independence is dropped, the bootstrap requires that the dependence be properly modelled.

Recall our original problem of estimating the variance of a sample mean. Only now, suppose that, unknown to the data analyst, the observations $(x_1, \ldots, x_n)$ are correlated. For example, assume that all of the observations have equal correlation ρ with one another

$$\mathrm{Cov}(x_i, x_j) \begin{array}{ll} = \sigma^2 & (i=j) \\ \\ = \rho\sigma^2 & (i\neq j) \end{array}$$

If the correlation is ignored, the introduction shows that the bootstrap estimate of the variance of $\bar{x}$ is $\Sigma(x_i-\bar{x})^2/n$. However, the actual variance of $\bar{x}$ is rather different: VAR $(\bar{x}) = \sigma^2\{1+\rho(n-1)\}/n \approx \rho\sigma^2$. A careless application of the bootstrap based on the wrong sampling procedure provides incorrect results. A multivariate example in which the obvious bootstrap approach fails appears in Beran and Srivastava (1985).

The bootstrap also gives misleading results for certain types of statistics. In general, such failures occur when the statistic of interest depends on a narrow feature of the original sampling process that bootstrap sampling cannot reproduce. For example, the bootstrap overcomes the well-publicized failure of the jackknife estimate of the standard error of the median, but fails for the maximum. Tukey (1987) gives a detailed heuristic argument, and a technical discussion appears in Bickel and Freedman (1981). Statistics such as the maximum that lack a normal sampling distribution require more caution than the usual weighed-average estimators so common in practice. Babu (1984) gives results for bootstrapping statistics that are asymptotically χ^2.

NOTES

1. The estimator var($\bar{x}$) can fail for a variety of reasons. Mosteller and Tukey (1977, chap. 7) point out the existence of other sources of variation, and the robustness literature (e.g., Hampel et al., 1986) contains many alternative estimators that perform better than s^2 if the population that has been sampled is not normal.

2. A finite total of n^n possible bootstrap samples exist, because any one of the n observations could be drawn first, any of the n could be second, and so forth. Not all of these samples give a distinct value for $\bar{x}^*$ because the mean ignores the ordering of the data. If we computed $\bar{x}^*$ for each of these n^n samples, we would obtain the true bootstrap variance of the sample mean, but such extreme computation is wasteful and unnecessary in this case.

3. The maximum likelihood estimator of σ^2 under a Gaussian population is $s_n{}^2$. Maximum likelihood estimators frequently are biased and lack corrections for degrees of freedom. Like maximum likelihood, the bootstrap typically leads to divisors of n rather than n−1. In a sense, the bootstrap *is* maximum likelihood, but with respect to the empirical distribution function.

4. The population defined by F_n is infinite in size, but only the observed set of values $(x_1, \ldots, x_n)$ are possible. Sampling with replacement from the observed data is equivalent to sampling from this infinite population.

5. With additional assumptions, one can obtain better estimates of the population distribution. For example, the parametric bootstrap uses an estimate of F that is a member of a particular parametric family, such as the Gaussian (Efron, 1982). The parametric bootstrap requires the rather strong assumption that we know the shape of the distribution of the population, and so we have chosen to stay with the basic scheme using F_n. The parametric approach does allow more detailed mathematical analysis of the technique.

6. The use of G^{*-1} to denote a quantile is standard in the statistics literature. This notation comes from recognizing that a quantile is really just a value of the inverse of a distribution function. Whereas a distribution function takes any value as an argument and returns a probability, the inverse of a distribution takes a probability and returns the associated quantile.

7. If the pairing given by sampling the z_i is removed, Y^* and X^* will be independent in the bootstrap simulation and $\hat{\beta}^*$ will be distributed about 0, with the exception of the constant. Nonparametric tests use this very idea. In these tests, all possible pairings (or a large sampling of pairings) of the regressors with the response are considered and the size of the observed effect is judged relative to this collection (see Lehmann and D'Abrera, 1975).

8. The residuals that are resampled must have an average of zero. If the residuals do not, as can occur when the regression model lacks a constant term, the bootstrap *fails* to give consistent variance estimates (Freedman, 1981).

9. Independence in the context of bootstrap resampling is always to be interpreted as conditional independence given the values of the observed data. This independence is a consequence of sampling with replacement from the original observations. The variance of the bootstrap population defined by the residuals is

$$\mathrm{VAR}(e_i{*}) = \sum_{i=1}^{n} e_i{}^2/n = \frac{n-k}{n}s^2$$

10. A linear statistic is a nonrandom linear combination of random variables. Thus, $\hat{\beta}$ is a linear statistic when the design is fixed. When the design is random, $\hat{\beta}$ is no longer linear because the weights of the linear combination vary with X.

11. The probability of not getting a particular observation in a bootstrap sample is the probability of choosing all n bootstrap observations from the remaining n−1 points, an event with probability $(1-1/n)^n \approx 0.36$.

12. The histograms of the bootstrap estimates have been smoothed using a *kernel density estimator*. The kernel density estimator smooths the random irregularities of the familiar histogram, removes some of the subjective choice of bin location and width, and gives a better visual impression of the shape of the distribution. Silverman (1986) gives an excellent overview of this technique. Further ideas on using kernel smoothing to improve bootstrap estimates appear in Silverman and Young (1987).

13. This naive illustrative model also suffers from specification error. The District of Columbia combines a high poverty percentage with a high average income, so that the poverty percentage may not be a good indicator of economic well-being.

14. An extensive discussion of problems caused by heteroscedasticity in regression appears in Carroll and Ruppert (1988). A more complex application of these ideas appears in Freedman and Peters (1982), who used the bootstrap to examine an econometric model that includes lagged endogenous variables. They found that the usual standard error was about one third of the true standard error. The bootstrap standard error was also too small, but much better than the usual WLS estimator (about 80% of the correct value).

15. That the robust estimator has smaller variation than the least-squares estimator does not contradict the Gauss-Markov theorem. This theorem only applies to linear statistics, and the robust estimator is *not* linear because its weights are determined iteratively from the data.

16. The notation for $t^*(\alpha;n)$ differs from the usual t-interval because the bootstrap is not making use of the notion of degrees of freedom, and only requires the sample size n.

17. As shown in the introduction, it is known that $\mathrm{VAR}^*\bar{x} = (n-1)s^2/n^2$. Thus, we could improve the bootstrap interval and reduce the calculations by making use of this fact and replace the bootstrap estimate sd_B^* by the true value, $(n-1)^{1/2}s/n$. In general, however, VAR* is seldom known and must be estimated by simulation. If one performs the simulations using $\sqrt{n}\,(\bar{x}^*-\bar{x})/s$, lacking s^* in the denominator, then the bootstrap interval does not give the desired coverage.

18. One can replace the normal distribution in assumption (A1) and those that follow by some other distribution (and $z(\alpha)$ by the correct percentile), but the large sample distribution of most statistics is Gaussian, and this choice is thus most broad. Also, these calculations are in terms of the true bootstrap distribution, not the simulated estimate of G_B^* from B replications. Generally, $B \approx 1000$ is necessary to get a good estimate of G^*.

19. The coverage of a prediction interval is the expected probability that the interval captures the future observation. Because the value being predicted is random, the situation differs from that with the usual confidence interval. A confidence interval either does or does not contain the sought parameter. A prediction interval, however, captures a fraction of the distribution of the predicted value. The average of this fraction over many samples is the coverage of the interval.

REFERENCES

AKRITAS, M. G. (1986) "Bootstrapping the Kaplan-Meier estimator." J. Amer. Stat. Assn. 81: 1032-1038.

BABU, G. J. (1984) "Bootstrapping statistics with linear combinations of chi-squares as weak limit." Sankhya, Series A, 46: 85-93.

BERAN, R., and SRIVASTAVA, M. S. (1985) "Bootstrap tests and confidence regions for functions of a covariance matrix." Annals of Statistics 13: 95-115.

BICKEL, P., and D. FREEDMAN (1981) "Some asymptotic theory for the bootstrap." Annals of Statistics 9: 1196-1217.

BISHOP, Y., S. FEINBURG, and P. HOLLAND (1975) Discrete Multivariate Analysis. Cambridge: MIT Press.

BOLLEN, K. A., and R. A. STINE (1988) "Bootstrapping indirect effects in structural equation models." (unpublished).

CARROLL, R. J., and D. RUPPERT (1988) "An asymptotic theory for weighted least-squares with weights estimated by replication." Biometrika 75: 35-44.

COOK, T. D., and D. T. CAMPBELL (1976) "The design and conduct of quasi-experiments and true experiments in field settings." In M. Dunnette, ed., Handbook of Industrial and Organization Psychology. New York: Rand-McNally.

DIACONIS, P., and B. EFRON (1983) "Computer intensive methods in statistics." Scientific American 248(5): 116-130.

DICICCO, T., and R. TIBSHIRANI (1987) "Bootstrap confidence intervals and bootstrap approximations." J. Amer. Stat. Assn. 82: 163-170.

EFRON, B. (1979a) "Computers and the theory of statistics: Thinking the unthinkable." Siam Rev. 21: 460-480.

EFRON, B. (1979b) "Bootstrap methods: another look at the jackknife." Annals of Statistics 7: 1-26.

EFRON, B. (1981) "Censored data and the bootstrap." J. Amer. Stat. Assn. 76: 312-319.

EFRON, B. (1982) The Jackknife, the Bootstrap, and Other Resampling Plans. CBMS 38, SIAM-NSF.

EFRON, B. (1983) "Estimating the error rate of a prediction rule: Improvement on cross-validation." J. Amer. Stat. Assn. 78: 316-331.

EFRON, B. (1987) "Better bootstrap confidence intervals."J. Amer. Stat. Assn. 82: 171-200.

EFRON, B., and G. GONG (1983) "A leisurely look at the bootstrap, the jackknife, and cross-validation." Amer. Statistician 37: 36-48.

EFRON, B., and R. TIBSHIRANI (1986) "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy." Stat. Science 1: 54-75.

FINIFTER, B. M. (1972) "The generation of confidence: Evaluating research findings by random subsample replication," pp. 112-175 in H. L. Costineau, ed., Sociological Methodology San Francisco: Jossey-Bass.

FOX, J. (1984) Linear Statistical Models and Related Methods. New York: Wiley.

FREEDMAN, D. A. (1981) "Bootstrapping regression models." Annals of Statistics 9: 1218-1228.

FREEDMAN, D. A. (1984) "On bootstrapping two-stage least squares estimates in stationary linear models." Annals of Statistics 12: 827-842.

FREEDMAN, D. A. and S. C. PETERS (1982) "Bootstrapping a regression equation: some empirical results." J. Amer. Stat. Assn. 79: 97-106.

GASSER, T., W. KOHLER, H. G. MULLER, A. KNEIP, R. LARJO, L. MOLINARI, and A. PRADER (1984) "Velocity and acceleration of height growth using kernel estimation." Annals of Human Biology 11: 397-411.

HALL, P. (1986a) "On the bootstrap and confidence intervals." Annals of Statistics 14: 1431-1452.

HALL, P. (1986b) "On the number of bootstrap simulations required to construct a confidence interval" Annals of Statistics 14: 1453-1462.

HALL, P. (1987) "On the bootstrap and likelihood-based confidence regions." Biometrika 74: 481-493.

HAMPEL, F. R., E. M. RONCHETTI, P. J. ROUSSEUW, and W. A. STAHEL (1986) Robust Statistics. New York: Wiley.

JORGENSEN, M. A. (1987) "Jackknifing fixed points of iterations." Biometrika 74; 207-211.

KATZ, S. H., M. L. HEDIGER, J. I. SCHALL, E. J. BOWERS, W. F. BARKER, S. AURAND, P. B. EVELETH, A. B. GRUSKIN, and J. S. PARKS (1980) "Blood pressure, growth, and maturation from childhood through adolescence." Hypertension 2 (Supp. 1): I55-I69.

LEHMANN, E. L., and H.J.M. D'ABRERA (1975) Nonparametrics. San Francisco: Holden-Day.

MILLER, R. G. (1974) "The jackknife—a review." Biometrika 61: 1-15.

MOSTELLER, F., and J. W. TUKEY (1977) Data Analysis and Regression. Reading, MA: Addison-Wesley.

RAO, J.N.K., and C.F.J. WU (1988) "Resampling inference with complex survey data." J. Amer. Stat. Assn. 83; 231-245.

SCHENKER, N. (1985) "Qualms about bootstrap confidence intervals." J. Amer. Stat. Assn. 80: 360-361.

SILVERMAN, B. W. (1985) "Some aspects of the spline smoothing approach to nonparametric regression curve fitting." J. Royal Stat. Society B47: 1-52.

SILVERMAN, B. W. (1986) Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.

SILVERMAN, B. W., and G. A. YOUNG (1987) "The bootstrap: To smooth or not to smooth?" Biometrika 74: 469-480.

SHORACK, G.R. (1982) "Bootstrapping robust regression." Comm. in Statistics A11: 961-972.

SOBEL, M. E. (1982) "Asymptotic confidence intervals for indirect effects in structural equation models," pp. 290-312 in H. L. Costineau, ed., Sociological Methodology. San Francisco, Jossey-Bass.

STINE, R. A. (1985) "Bootstrap prediction intervals for regression." J. Amer. Stat. Assn. 80: 1026-1031.

STINE, R. A. (1987) "Estimating properties of autoregressive forecasts." J. Amer. Stat. Assn. 82: 1072-1078.

STREET, J. O., R. J. CARROLL, and D. RUPPERT (1988) "A note on computing robust regression estimates via iteratively reweighted least squares." Amer. Statistician 42: 152-154.

SWANEPOEL, J.W.H., and J.W.J. VAN WYK (1986) "The bootstrap applied to power spectral density function estimation." Biometrika 73: 135-142.

TIBSHIRANI, R. (1985) "How many bootstraps?" Technical report no. 362, Dept. of Statistics, Stanford University.

TIBSHIRANI, R. (1988) "Variance stabilization and the bootstrap." Biometrika 75: 433-444.

TUKEY, J. W. (1958) "Bias and confidence in not-quite large samples." Annals of Math. Statistics 29: 614.

TUKEY, J. W. (1986) "Sunset salvo." Amer. Statistician 40: 72-76.

TUKEY, J. W. (1987) "Kinds of bootstraps and kinds of jackknives, discussed in terms of a year of weather related data." Technical report no. 292, Dept. of Statistics, Princeton University.

WU, C.F.J. (1986) "Jackknife, bootstrap and other resampling methods in regression analysis." Annals of Statistics 14; 1261-1350.

ZACKS, S. (1971) The Theory of Statistical Inference. New York: Wiley.

Robert Stine is Associate Professor of Statistics at The Wharton School of the University of Pennsylvania. His Ph.D. is in statistics from Princeton University and his dissertation proposed the use of bootstrap methods for obtaining better prediction intervals in forecasting problems. His current research interests include resampling methods, time-series analysis, and statistical computing. In each case, his work addresses problems that arise in data analysis, particularly the analysis of data related to energy resources and consumption and to biomedical problems. Dr. Stine has also given a lecture series on bootstrap methods at the Summer Program of the Inter-University Consortium for Political and Social Research.