21 Bootstrapping Regression Models

Bootstrapping is a nonparametric approach to statistical inference that substitutes computation for more traditional distributional assumptions and asymptotic results.[1] Bootstrapping offers a number of advantages:

- The bootstrap is quite general, although there are some cases in which it fails.
- Because it does not require distributional assumptions (such as normally distributed errors), the bootstrap can provide more accurate inferences when the data are not well behaved or when the sample size is small.
- It is possible to apply the bootstrap to statistics with sampling distributions that are difficult to derive, even asymptotically.
- It is relatively simple to apply the bootstrap to complex data-collection plans (such as stratified and clustered samples).

21.1 Bootstrapping Basics

My principal aim is to explain how to bootstrap regression models (broadly construed to include generalized linear models, etc.), but the topic is best introduced in a simpler context: Suppose that we draw an independent random sample from a large population.[2] For concreteness and simplicity, imagine that we sample four working, married couples, determining in each case the husband's and wife's income, as recorded in Table 21.1. I will focus on the difference in incomes between husbands and wives, denoted as Y_i for the ith couple.

We want to estimate the mean difference in income between husbands and wives in the population. Please bear with me as I review some basic statistical theory: A point estimate of this population mean difference μ is the sample mean,

$$\overline{Y} = \sum \frac{Y_i}{n} = \frac{6 - 3 + 5 + 3}{4} = 2.75$$

Elementary statistical theory tells us that the standard deviation of the sampling distribution of sample means is $\mathrm{SD}(\overline{Y}) = \sigma/\sqrt{n}$, where σ is the population standard deviation of Y.

If we knew σ, and if Y were normally distributed, then a 95% confidence interval for μ would be

$$\mu = \overline{Y} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

[1] The term *bootstrapping*, coined by Efron (1979), refers to using the sample to learn about the sampling distribution of a statistic without reference to external assumptions—as in "pulling oneself up by one's bootstraps."

[2] In an *independent random sample*, each element of the population can be selected more than once. In a *simple random sample*, in contrast, once an element is selected into the sample, it is removed from the population, so that sampling is done "without replacement." When the population is very large in comparison to the sample (say, at least 10 times as large), the distinction between independent and simple random sampling becomes inconsequential.

Table 21.1 Contrived "Sample" of Four Married Couples, Showing Husbands' and Wives' Incomes in Thousands of Dollars

Observation	*Husband's Income*	*Wife's Income*	*Difference* Y_i
1	24	18	6
2	14	17	−3
3	40	35	5
4	44	41	3

where $z_{.025} = 1.96$ is the standard normal value with a probability of .025 to the right. If Y is *not* normally distributed in the population, then this result applies asymptotically. Of course, the asymptotics are cold comfort when $n = 4$.

In a real application, we do not know σ. The standard estimator of σ is

$$S = \sqrt{\frac{\sum(Y_i - \overline{Y})^2}{n-1}}$$

from which the standard error of the mean (i.e., the *estimated* standard deviation of $\overline{Y}$) is $\text{SE}(\overline{Y}) = S/\sqrt{n}$. If the population is normally distributed, then we can take account of the added uncertainty associated with estimating the standard deviation of the mean by substituting the heavier-tailed t-distribution for the normal distribution, producing the 95% confidence interval

$$\mu = \overline{Y} \pm t_{n-1,\,.025}\frac{S}{\sqrt{n}}$$

Here, $t_{n-1,\,.025}$ is the critical value of t with $n-1$ degrees of freedom and a right-tail probability of .025.

In the present case, $S = 4.031$, $\text{SE}(\overline{Y}) = 4.031/\sqrt{4} = 2.015$, and $t_{3,\,.025} = 3.182$. The 95% confidence interval for the population mean is thus

$$\mu = 2.75 \pm 3.182 \times 2.015 = 2.75 \pm 6.41$$

or, equivalently,

$$-3.66 < \mu < 9.16$$

As one would expect, this confidence interval—which is based on only four observations—is very wide and includes 0. It is, unfortunately, hard to be sure that the population is reasonably close to normally distributed when we have such a small sample, and so the t-interval may not be valid.[3]

Bootstrapping begins by using the distribution of data values in the sample (here, $Y_1 = 6$, $Y_2 = -3$, $Y_3 = 5$, $Y_4 = 3$) to *estimate* the distribution of Y in the population.[4] That is, we define the random variable Y^* with distribution[5]

[3] To say that a confidence interval is "valid" means that it has the stated coverage. That is, a 95% confidence interval is valid if it is constructed according to a procedure that encloses the population mean in 95% of samples.

[4] An alternative would be to resample from a distribution given by a nonparametric density estimate (see, e.g., Silverman & Young, 1987). Typically, however, little if anything is gained by using a more complex estimate of the population distribution. Moreover, the simpler method explained here generalizes more readily to more complex situations in which the population is multivariate or not simply characterized by a distribution.

[5] The asterisks on $p^*(\cdot)$, E^*, and V^* remind us that this probability distribution, expectation, and variance are conditional on the specific sample in hand. Were we to select another sample, the values of Y_1, Y_2, Y_3, and Y_4, would change and—along with them—the probability distribution of Y^*, its expectation, and variance.

y^*	$p^*(y^*)$
6	.25
−3	.25
5	.25
3	.25

Note that

$$E^*(Y^*) = \sum_{\text{all } y^*} y^* p(y^*) = 2.75 = \overline{Y}$$

and

$$\begin{aligned} V^*(Y^*) &= \sum [y^* - E^*(Y^*)]^2 p(y^*) \\ &= 12.187 = \frac{3}{4} S^2 = \frac{n-1}{n} S^2 \end{aligned}$$

Thus, the expectation of Y^* is just the sample mean of Y, and the variance of Y^* is [except for the factor $(n-1)/n$, which is trivial in larger samples] the sample variance of Y.

We next mimic sampling from the original population by treating the sample as if it were the population, enumerating all possible samples of size $n = 4$ from the probability distribution of Y^*. In the present case, each *bootstrap sample* selects four values *with replacement* from among the four values of the original sample. There are, therefore, $4^4 = 256$ different bootstrap samples,[6] each selected with probability 1/256. A few of the 256 samples are shown in Table 21.2. Because the four observations in each bootstrap sample are chosen with replacement, particular bootstrap samples usually have repeated observations from the original sample. Indeed, of the illustrative bootstrap samples shown in Table 21.2, only sample 100 does *not* have repeated observations.

Let us denote the bth bootstrap sample[7] as $\mathbf{y}_b^* = [Y_{b1}^*, Y_{b2}^*, Y_{b3}^*, Y_{b4}^*]'$, or more generally, $\mathbf{y}_b^* = [Y_{b1}^*, Y_{b2}^*, \ldots, Y_{bn}^*]'$, where $b = 1, 2, \ldots, n^n$. For each such bootstrap sample, we calculate the mean,

$$\overline{Y}_b^* = \frac{\sum_{i=1}^n Y_{bi}^*}{n}$$

The sampling distribution of the 256 bootstrap means is shown in Figure 21.1.

The mean of the 256 bootstrap sample means is just the original sample mean, $\overline{Y} = 2.75$. The standard deviation of the bootstrap means is

$$\begin{aligned} \text{SD}^*(\overline{Y}^*) &= \sqrt{\frac{\sum_{b=1}^{n^n} (\overline{Y}_b^* - \overline{Y})^2}{n^n}} \\ &= 1.745 \end{aligned}$$

We divide here by n^n rather than by $n^n - 1$ because the distribution of the $n^n = 256$ bootstrap sample means (Figure 21.1) is known, *not* estimated. The standard deviation of the bootstrap

[6] Many of the 256 samples have the same elements but in different order—for example, [6, 3, 5, 3] and [3, 5, 6, 3]. We could enumerate the unique samples without respect to order and find the probability of each, but it is simpler to work with the 256 orderings because each ordering has equal probability.

[7] If vector notation is unfamiliar, then think of $\mathbf{y}_b^*$ simply as a list of the bootstrap observations Y_{bi}^* for sample b.

Table 21.2 A Few of the 256 Bootstrap Samples for the Data Set [6, −3, 5, 3], and the Corresponding Bootstrap Means, $\overline{Y}^*_b$

Bootstrap Sample b	Y^*_{b1}	Y^*_{b2}	Y^*_{b3}	Y^*_{b4}	$\overline{Y}^*_b$
1	6	6	6	6	6.00
2	6	6	6	−3	3.75
3	6	6	6	5	5.75
⋮	⋮				⋮
100	−3	5	6	3	2.75
101	−3	5	−3	6	1.25
⋮	⋮				⋮
255	3	3	3	5	3.50
256	3	3	3	3	3.00

means is nearly equal to the usual standard error of the sample mean; the slight slippage is due to the factor $\sqrt{n/(n-1)}$, which is typically negligible (though not when $n = 4$):[8]

$$\text{SE}(\overline{Y}) = \sqrt{\frac{n}{n-1}}\text{SD}^*(\overline{Y}^*)$$

$$2.015 = \sqrt{\frac{4}{3}} \times 1.745$$

This precise relationship between the usual formula for the standard error and the bootstrap standard deviation is peculiar to *linear statistics* (i.e., linear functions of the data) like the mean. For the mean, then, the bootstrap standard deviation is just a more complicated way to calculate what we already know, but

- bootstrapping might still provide more accurate confidence intervals, as I will explain presently; and
- bootstrapping can be applied to *nonlinear* statistics for which we do not have standard-error formulas or for which only asymptotic standard errors are available.

Bootstrapping exploits the following central analogy:

The population is to the sample
as
the sample is to the bootstrap samples.

Consequently,

- the *bootstrap observations* Y^*_{bi} are analogous to the *original observations* Y_i;
- the *bootstrap mean* $\overline{Y}^*_b$ is analogous to the *mean of the original sample* $\overline{Y}$;
- the *mean of the original sample* $\overline{Y}$ is analogous to the (unknown) *population mean* μ; and
- the *distribution of the bootstrap sample means* is analogous to the (unknown) *sampling distribution of means* for samples of size n drawn from the original population.

[8] See Exercise 21.1.

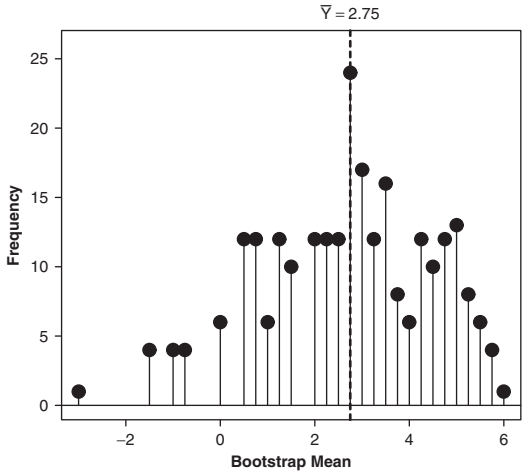

Figure 21.1 Graph of the 256 bootstrap means from the sample [6, −3, 5, 3]. The broken vertical line gives the mean of the original sample, $\overline{Y} = 2.75$, which is also the mean of the 256 bootstrap means.

> Bootstrapping uses the sample data to estimate relevant characteristics of the population. The sampling distribution of a statistic is then constructed empirically by resampling from the sample. The resampling procedure is designed to parallel the process by which sample observations were drawn from the population. For example, if the data represent an independent random sample of size n (or a simple random sample of size n from a much larger population), then each bootstrap sample selects n observations with replacement from the original sample. The key bootstrap analogy is the following: *The population is to the sample as the sample is to the bootstrap samples*.

The bootstrapping calculations that we have undertaken thus far depend on very small sample size, because the number of bootstrap samples (n^n) quickly becomes unmanageable: Even for samples as small as $n = 10$, it is impractical to enumerate all the $10^{10} = 10$ billion bootstrap samples. Consider the "data" shown in Table 21.3, an extension of the previous example. The mean and standard deviation of the differences in income Y are $\overline{Y} = 4.6$ and $S = 5.948$. Thus, the standard error of the sample mean is $\text{SE}(\overline{Y}) = 5.948/\sqrt{10} = 1.881$.

Although we cannot (as a practical matter) enumerate *all* the 10^{10} bootstrap samples, it is easy to draw at random a large number of bootstrap samples. To estimate the standard deviation of a statistic (here, the mean)—that is, to get a bootstrap standard error—100 or 200 bootstrap samples should be more than sufficient. To find a confidence interval, we will need a larger number of bootstrap samples, say 1,000 or 2,000.[9]

A practical bootstrapping procedure, therefore, is as follows:

1. Let r denote the number of *bootstrap replications*—that is, the number of bootstrap samples to be selected.

[9]Results presented by Efron and Tibshirani (1993, chap. 19) suggest that basing bootstrap confidence intervals on 1,000 bootstrap samples generally provides accurate results, and using 2,000 bootstrap replications should be very safe.

Table 21.3 Contrived "Sample" of 10 Married Couples, Showing Husbands' and Wives' Incomes in Thousands of Dollars

Observation	*Husband's Income*	*Wife's Income*	*Difference* Y_i
1	24	18	6
2	14	17	−3
3	40	35	5
4	44	41	3
5	24	18	6
6	19	9	10
7	21	10	11
8	22	30	−8
9	30	23	7
10	24	15	9

2. For each bootstrap sample $b = 1, \ldots, r$, randomly draw n observations $Y^*_{b1}, Y^*_{b2}, \ldots, Y^*_{bn}$ with replacement from among the n sample values, and calculate the bootstrap sample mean,

$$\overline{Y}^*_b = \frac{\sum_{i=1}^n Y^*_{bi}}{n}$$

3. From the r bootstrap samples, *estimate* the standard deviation of the bootstrap means:[10]

$$\mathrm{SE}^*(\overline{Y}^*) = \sqrt{\frac{\sum_{b=1}^r \left(\overline{Y}^*_b - \overline{\overline{Y}}^*\right)^2}{r-1}}$$

where

$$\overline{\overline{Y}}^* \equiv \frac{\sum_{b=1}^r \overline{Y}^*_b}{r}$$

is the mean of the bootstrap means. We can, if we wish, "correct" $\mathrm{SE}^*(\overline{Y}^*)$ for degrees of freedom, multiplying by $\sqrt{n/(n-1)}$.

To illustrate this procedure, I drew $r = 2,000$ bootstrap samples, each of size $n = 10$, from the "data" given in Table 21.3, calculating the mean, $\overline{Y}^*_b$, for each sample. A few of the 2,000 bootstrap replications are shown in Table 21.4, and the distribution of bootstrap means is graphed in Figure 21.2.

We know from statistical theory that were we to enumerate all the 10^{10} bootstrap samples (or, alternatively, to sample infinitely from the population of bootstrap samples), the average bootstrap mean would be $E^*(\overline{Y}^*) = \overline{Y} = 4.6$, and the standard deviation of the bootstrap means would be

[10] It is important to distinguish between the "ideal" bootstrap estimate of the standard deviation of the mean, $\mathrm{SD}^*(\overline{Y}^*)$, which is based on *all* n^n bootstrap samples, and the *estimate* of this quantity, $\mathrm{SE}^*(\overline{Y}^*)$, which is based on r randomly selected bootstrap samples. By making r large enough, we seek to ensure that $\mathrm{SE}^*(\overline{Y}^*)$ is close to $\mathrm{SD}^*(\overline{Y}^*)$. Even $\mathrm{SD}^*(\overline{Y}^*) = \mathrm{SE}(\overline{Y})$ is an imperfect estimate of the true standard deviation of the sample mean $\mathrm{SD}(\overline{Y})$, however, because it is based on a *particular sample* of size n drawn from the original population.

Table 21.4 A Few of the $r = 2,000$ Bootstrap Samples Drawn From the Data Set [6, −3, 5, 3, 6, 10, 11, −8, 7, 9] and the Corresponding Bootstrap Means, $\overline{Y}^*_b$

b	Y^*_{b1}	Y^*_{b2}	Y^*_{b3}	Y^*_{b4}	Y^*_{b5}	Y^*_{b6}	Y^*_{b7}	Y^*_{b8}	Y^*_{b9}	Y^*_{b10}	$\overline{Y}^*_b$
1	6	10	6	5	−8	9	9	6	11	3	5.7
2	9	9	7	7	3	3	−3	−3	−8	6	3.0
3	9	−3	6	5	10	6	10	10	10	6	6.9
⋮	⋮										⋮
1999	6	9	6	3	11	6	6	7	3	9	6.6
2000	7	6	7	3	10	6	9	3	10	6	6.7

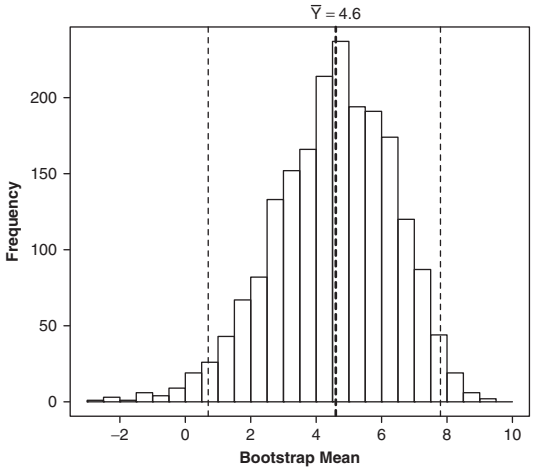

Figure 21.2 Histogram of $r = 2{,}000$ bootstrap means, produced by resampling from the "sample" [6, −3, 5, 3, 6, 10, 11, −8, 7, 9]. The heavier broken vertical line gives the sample mean, $\overline{Y} = 4.6$; the lighter broken vertical lines give the boundaries of the 95% percentile confidence interval for the population mean μ based on the 2,000 bootstrap samples. The procedure for constructing this confidence interval is described in the next section.

$\text{SE}^*(\overline{Y}^*) = \text{SE}(\overline{Y})\sqrt{(n-1)/n} = 1.881\sqrt{9/10} = 1.784$. For the 2,000 bootstrap samples that I selected, $\overline{\overline{Y}}^* = 4.693$ and $\text{SE}(\overline{Y}^*) = 1.750$—both quite close to the theoretical values.

The bootstrapping procedure described in this section can be generalized to derive the empirical sampling distribution for an estimator $\widehat{\theta}$ of the parameter θ:

1. Specify the data-collection scheme $\mathcal{S}$ that gives rise to the observed sample when applied to the population:[11]

$$\mathcal{S}(\text{Population}) \Longrightarrow \text{Sample}$$

[11]The "population" can be real—the population of working married couples—or hypothetical—the population of conceivable replications of an experiment. What is important in the present context is that the sampling procedure can be described concretely.

The estimator $\widehat{\theta}$ is some function $s(\cdot)$ of the observed sample. In the preceding example, the data-collection procedure is independent random sampling from a large population.

2. Using the observed sample data as a "stand-in" for the population, replicate the data-collection procedure, producing r bootstrap samples:

$$\mathcal{S}(\text{Sample}) \left\{ \begin{array}{ll} \Longrightarrow & \text{Bootstrap sample}_1 \\ \Longrightarrow & \text{Bootstrap sample}_2 \\ & \vdots \\ \Longrightarrow & \text{Bootstrap sample}_r \end{array} \right.$$

3. For each bootstrap sample, calculate the estimate $\widehat{\theta}^*_b = s(\text{Bootstrap sample}_b)$.

4. Use the distribution of the $\widehat{\theta}^*_b$s to estimate properties of the sampling distribution of $\widehat{\theta}$. For example, the bootstrap standard error of $\widehat{\theta}$ is $\text{SE}^*(\widehat{\theta}^*)$ (i.e., the standard deviation of the r bootstrap replications $\widehat{\theta}^*_b$):[12]

$$\text{SE}^*(\widehat{\theta}^*) \equiv \sqrt{\frac{\sum_{b=1}^{r}(\widehat{\theta}^*_b - \overline{\theta}^*)^2}{r-1}}$$

where

$$\overline{\theta}^* \equiv \frac{\sum_{b=1}^{r}\widehat{\theta}^*_b}{r}$$

21.2 Bootstrap Confidence Intervals

21.2.1 Normal-Theory Intervals

As I have mentioned, normal-theory confidence intervals for means are based on the t-distribution when the population variance of Y is unknown. Most statistics, including sample means, are asymptotically normally distributed; in large samples we can therefore use the bootstrap standard error, along with the normal distribution, to produce a $100(1-a)\%$ confidence interval for θ based on the estimator $\widehat{\theta}$:

$$\theta = \widehat{\theta} \pm z_{a/2}\text{SE}^*(\widehat{\theta}^*) \tag{21.1}$$

In Equation 21.1, $z_{a/2}$ is the standard normal value with probability $a/2$ to the right. This approach will work well if the bootstrap sampling distribution of the estimator is approximately normal, and so it is advisable to examine a normal quantile-comparison plot of the bootstrap distribution.

There is no advantage to calculating normal-theory bootstrap confidence intervals for linear statistics like the mean, because in this case the ideal bootstrap standard deviation of the statistic and the standard error based directly on the sample coincide. Using bootstrap resampling in this setting just makes for extra work and introduces an additional small random component into standard errors.

[12] We may want to apply the correction factor $\sqrt{n/(n-1)}$.

Having produced r bootstrap replicates $\widehat{\theta}^*_b$ of an estimator $\widehat{\theta}$, the bootstrap standard error is the standard deviation of the bootstrap replicates: $\text{SE}^*(\widehat{\theta}^*) = \sqrt{\sum_{b=1}^{r}(\widehat{\theta}^*_b - \overline{\theta}^*)^2/(r-1)}$, where $\overline{\theta}^*$ is the mean of the $\widehat{\theta}^*_b$. In large samples, where we can rely on the normality of $\widehat{\theta}$, a 95% confidence interval for θ is given by $\widehat{\theta} \pm 1.96\ \text{SE}^*(\widehat{\theta}^*)$.

21.2.2 Percentile Intervals

Another very simple approach is to use the quantiles of the bootstrap sampling distribution of the estimator to establish the end points of a confidence interval *nonparametrically*. Let $\widehat{\theta}^*_{(b)}$ represent the ordered bootstrap estimates, and suppose that we want to construct a $(100 - a)\%$ confidence interval. If the number of bootstrap replications r is large (as it should be to construct a percentile interval), then the $a/2$ and $1 - a/2$ quantiles of $\widehat{\theta}^*_b$ are approximately $\widehat{\theta}^*_{(\text{lower})}$ and $\widehat{\theta}^*_{(\text{upper})}$, where lower $= ra/2$ and upper $= r(1 - a/2)$. If lower and upper are not integers, then we can interpolate between adjacent ordered values $\widehat{\theta}^*_{(b)}$ or round off to the nearest integer.

A nonparametric confidence interval for θ can be constructed from the quantiles of the bootstrap sampling distribution of $\widehat{\theta}^*$. The 95% percentile interval is $\widehat{\theta}^*_{(\text{lower})} < \theta < \widehat{\theta}^*_{(\text{upper})}$, where the $\widehat{\theta}^*_{(b)}$ are the r ordered bootstrap replicates; lower $= .025 \times r$ and upper $= .975 \times r$.

A 95% confidence interval for the $r = 2,000$ resampled means in Figure 21.2, for example, is constructed as follows:

$$
\begin{aligned}
\text{lower} &= 2000(.05/2) = 50 \\
\text{upper} &= 2000(1 - .05/2) = 1950 \\
\overline{Y}^*_{(50)} &= 0.7 \\
\overline{Y}^*_{(1950)} &= 7.8 \\
0.7 < \mu &< 7.8
\end{aligned}
$$

The end points of this interval are marked in Figure 21.2. Because of the skew of the bootstrap distribution, the percentile interval is not quite symmetric around $\overline{Y} = 4.6$. By way of comparison, the standard t-interval for the mean of the original sample of 10 observations is

$$
\begin{aligned}
\mu &= \overline{Y} \pm t_{9,\,.025}\text{SE}(\overline{Y}) \\
&= 4.6 \pm 2.262 \times 1.881 \\
&= 4.6 \pm 4.255 \\
0.345 < \mu &< 8.855
\end{aligned}
$$

In this case, the standard interval is a bit wider than the percentile interval, especially at the top.

21.2.3 Improved Bootstrap Intervals

I will briefly describe an adjustment to percentile intervals that improves their accuracy.[13] As before, we want to produce a $100(1-a)\%$ confidence interval for θ having computed the sample estimate $\widehat{\theta}$ and bootstrap replicates $\widehat{\theta}^*_b$; $b = 1, \ldots, r$. We require $z_{a/2}$, the unit-normal value with probability $a/2$ to the right, and two "correction factors," Z and A, defined in the following manner:

- Calculate

$$Z \equiv \Phi^{-1}\left[\frac{\#_{b=1}^{r}(\widehat{\theta}^*_b < \widehat{\theta})}{r}\right]$$

  where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function, and $\#(\widehat{\theta}^*_b < \widehat{\theta})/r$ is the proportion of bootstrap replicates below the estimate $\widehat{\theta}$. If the bootstrap sampling distribution is symmetric and if $\widehat{\theta}$ is unbiased, then this proportion will be close to .5, and the "correction factor" Z will be close to 0.
- Let $\widehat{\theta}_{(-i)}$ represent the value of $\widehat{\theta}$ produced when the ith observation is deleted from the sample;[14] there are n of these quantities. Let $\overline{\theta}$ represent the average of the $\widehat{\theta}_{(-i)}$; that is, $\overline{\theta} \equiv \sum_{i=1}^{n} \widehat{\theta}_{(-i)}/n$. Then calculate

$$A \equiv \frac{\sum_{i=1}^{n}(\widehat{\theta}_{(-i)} - \overline{\theta})^3}{6[\sum_{i=1}^{n}(\widehat{\theta}_{(-i)} - \overline{\theta})^2]^{3/2}} \tag{21.2}$$

With the correction factors Z and A in hand, compute

$$A_1 \equiv \Phi\left[Z + \frac{Z - z_{a/2}}{1 - A(Z - z_{a/2})}\right]$$
$$A_2 \equiv \Phi\left[Z + \frac{Z + z_{a/2}}{1 - A(Z + z_{a/2})}\right]$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution function. Note that when the correction factors Z and A are both 0, $A_1 = \Phi(-z_{a/2}) = a/2$, and $A_2 = \Phi(z_{a/2}) = 1 - a/2$. The values A_1 and A_2 are used to locate the end points of the corrected percentile confidence interval. In particular, the corrected interval is

$$\widehat{\theta}^*_{(\text{lower}^*)} < \theta < \widehat{\theta}^*_{(\text{upper}^*)}$$

where lower* $= rA_1$ and upper* $= rA_2$ (rounding or interpolating as required).

> The lower and upper bounds of percentile confidence intervals can be corrected to improve the accuracy of these intervals.

[13] The interval described here is called a "bias-corrected, accelerated" (or BC$_a$) percentile interval. Details can be found in Efron and Tibshirani (1993, chap. 14); also see Stine (1990) for a discussion of different procedures for constructing bootstrap confidence intervals.

[14] The $\widehat{\theta}_{(-i)}$ are called the *jackknife values* of the statistic $\widehat{\theta}$. The jackknife values can also be used as an alternative to the bootstrap to find a nonparametric confidence interval for θ. See Exercise 21.2.

Applying this procedure to the "data" in Table 21.3, we have $z_{.05/2} = 1.96$ for a 95% confidence interval. There are 926 bootstrapped means below $\overline{Y} = 4.6$, and so $Z = \Phi^{-1}(926/2000) = -0.09288$. The $\overline{Y}_{(-i)}$ are 4.444, 5.444, ..., 4.111; the mean of these values is $\overline{\overline{Y}} = \overline{Y} = 4.6$,[15] and (from Equation 21.2) $A = -0.05630$. Using these correction factors,

$$\begin{aligned} A_1 &= \Phi\left\{-0.09288 + \frac{-0.09288 - 1.96}{1 - [-.05630(-0.09288 - 1.96)]}\right\} \\ &= \Phi(-2.414) = 0.007889 \\ A_2 &= \Phi\left\{-0.09288 + \frac{-0.09288 + 1.96}{1 - [-.05630(-0.09288 + 1.96)]}\right\} \\ &= \Phi(1.597) = 0.9449 \end{aligned}$$

Multiplying by r, we have $2000 \times .0.007889 \approx 16$ and $2000 \times .0.9449 \approx 1890$, from which

$$\begin{aligned} \overline{Y}^*_{(16)} < \mu < \overline{Y}^*_{(1890)} \\ -0.4 < \mu < 7.3 \end{aligned} \tag{21.3}$$

Unlike the other confidence intervals that we have calculated for the "sample" of 10 differences in income between husbands and wives, the interval given in Equation 21.3 includes 0.

21.3 Bootstrapping Regression Models

The procedures of the previous section can be easily extended to regression models. The most straightforward approach is to collect the response-variable value and regressors for each observation

$$\mathbf{z}_i' \equiv [Y_i, X_{i1}, \ldots, X_{ik}]$$

Then the observations $\mathbf{z}_1', \mathbf{z}_2', \ldots, \mathbf{z}_n'$ can be resampled, and the regression estimator computed for each of the resulting bootstrap samples, $\mathbf{z}^{*\prime}_{b1}, \mathbf{z}^{*\prime}_{b2}, \ldots, \mathbf{z}^{*\prime}_{bn}$, producing r sets of bootstrap regression coefficients, $\mathbf{b}^*_b = [A^*_b, B^*_{b1}, \ldots, B^*_{bk}]'$. The methods of the previous section can be applied to compute standard errors or confidence intervals for the regression estimates.

Directly resampling the observations $\mathbf{z}_i'$ implicitly treats the regressors $X_1, \ldots, X_k$ as *random* rather than *fixed*. We may want to treat the Xs as fixed (if, e.g., the data derive from an experimental design). In the case of linear regression, for example,

1. Estimate the regression coefficients $A, B_1, \ldots, B_k$ for the original sample, and calculate the fitted value and residual for each observation:

$$\begin{aligned} \widehat{Y}_i &= A + B_1 x_{i1} + \cdots + B_k x_{ik} \\ E_i &= Y_i - \widehat{Y}_i \end{aligned}$$

2. Select bootstrap samples of the *residuals*, $\mathbf{e}^*_b = [E^*_{b1}, E^*_{b2}, \ldots, E^*_{bn}]'$, and from these, calculate bootstrapped Y values, $\mathbf{y}^*_b = [Y^*_{b1}, Y^*_{b2}, \ldots, Y^*_{bn}]'$, where $Y^*_{bi} = \widehat{Y}_i + E^*_{bi}$.

[15] The average of the jackknifed estimates is not, in general, the same as the estimate calculated for the full sample, but this *is* the case for the jackknifed sample means. See Exercise 21.2.

3. Regress the bootstrapped Y values on the *fixed* X values to obtain bootstrap regression coefficients.
   *If, for example, estimates are calculated by least-squares regression, then $\mathbf{b}_b^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_b^*$ for $b = 1, \ldots, r$.
4. The resampled $\mathbf{b}_b^* = [A_b^*, B_{b1}^*, \ldots, B_{bk}^*]'$ can be used in the usual manner to construct bootstrap standard errors and confidence intervals for the regression coefficients.

Bootstrapping with fixed X draws an analogy between the fitted value $\widehat{Y}$ in the sample and the conditional expectation of Y in the population, and between the residual E in the sample and the error ε in the population. Although no assumption is made about the *shape* of the error distribution, the bootstrapping procedure, by constructing the Y_{bi}^* according to the linear model, implicitly assumes that the functional form of the model is correct.

Furthermore, by resampling residuals and randomly reattaching them to fitted values, the procedure implicitly assumes that the errors are *identically distributed*. If, for example, the true errors have nonconstant variance, then this property will *not* be reflected in the resampled residuals. Likewise, the unique impact of a high-leverage outlier will be lost to the resampling.[16]

> Regression models and similar statistical models can be bootstrapped by (1) treating the regressors as random and selecting bootstrap samples directly from the observations $\mathbf{z}_i' = [Y_i, X_{i1}, \ldots, X_{ik}]$ or (2) treating the regressors as fixed and resampling from the residuals E_i of the fitted regression model. In the latter instance, bootstrap observations are constructed as $Y_{bi}^* = \widehat{Y}_i + E_{bi}^*$, where the $\widehat{Y}_i$ are the fitted values from the original regression, and the E_{bi}^* are the resampled residuals for the bth bootstrap sample. In each bootstrap sample, the Y_{bi}^* are then regressed on the original Xs. A disadvantage of fixed-X resampling is that the procedure implicitly assumes that the functional form of the regression model fit to the data is correct and that the errors are identically distributed.

To illustrate bootstrapping regression coefficients, I will use Duncan's regression of occupational prestige on the income and educational levels of 45 U.S. occupations.[17] The Huber M estimator applied to Duncan's regression produces the following fit, with asymptotic standard errors shown in parentheses beneath each coefficient:[18]

$$\begin{array}{rcrcrcr} \widehat{\text{Prestige}} & = & -7.289 & + & 0.7104 & \text{Income} & + \quad 0.4819 \quad \text{Education} \\ & & (3.588) & & (0.1005) & & (0.0825) \end{array}$$

Using random resampling, I drew $r = 2,000$ bootstrap samples, calculating the Huber estimator for each bootstrap sample. The results of this computationally intensive procedure are summarized in Table 21.5. The distributions of the bootstrapped regression coefficients for income and education are graphed in Figure 21.3(a) and (b), along with the percentile confidence intervals for these coefficients. Figure 21.3(c) shows a scatterplot of the bootstrapped coefficients

[16] For these reasons, random-X resampling may be preferable even if the X values are best conceived as fixed. See Exercise 21.3.

[17] These data were discussed in Chapter 19 on robust regression and at several other points in this text.

[18] M estimation is a method of robust regression described in Section 19.1.

Table 21.5 Statistics for $r = 2,000$ Bootstrapped Huber Regressions Applied to Duncan's Occupational Prestige Data

	Coefficient		
	Constant	*Income*	*Education*
Average bootstrap estimate	−7.001	0.6903	0.4918
Bootstrap standard error	3.165	0.1798	0.1417
Asymptotic standard error	3.588	0.1005	0.0825
Normal-theory interval	(−13.423,−1.018)	(0.3603,1.0650)	(0.2013,0.7569)
Percentile interval	(−13.150,−0.577)	(0.3205,1.0331)	(0.2030,0.7852)
Adjusted percentile interval	(−12.935,−0.361)	(0.2421,0.9575)	(0.2511,0.8356)

NOTES: Three bootstrap confidence intervals are shown for each coefficient. Asymptotic standard errors are also shown for comparison.

for income and education, which gives a sense of the covariation of the two estimates; it is clear that the income and education coefficients are strongly negatively correlated.[19]

The bootstrap standard errors of the income and education coefficients are much larger than the asymptotic standard errors, underscoring the inadequacy of the latter in small samples. The simple normal-theory confidence intervals based on the bootstrap standard errors (and formed as the estimated coefficients ± 1.96 standard errors) are reasonably similar to the percentile intervals for the income and education coefficients; the percentile intervals differ slightly from the adjusted percentile intervals. Comparing the average bootstrap coefficients $\overline{A}^*$, $\overline{B}_1^*$, and $\overline{B}_2^*$ with the corresponding estimates A, B_1, and B_2 suggests that there is little, if any, bias in the Huber estimates.[20]

21.4 Bootstrap Hypothesis Tests*

In addition to providing standard errors and confidence intervals, the bootstrap can also be used to test statistical hypotheses. The application of the bootstrap to hypothesis testing is more or less obvious for individual coefficients because a bootstrap confidence interval can be used to test the hypothesis that the corresponding parameter is equal to any specific value (typically 0 for a regression coefficient).

More generally, let $T \equiv t(\mathbf{z})$ represent a test statistic, written as a function of the sample $\mathbf{z}$. The contents of $\mathbf{z}$ vary by context. In regression analysis, for example, $\mathbf{z}$ is the $n \times k + 1$ matrix $[\mathbf{y}, \mathbf{X}]$ containing the response variable and the regressors.

For concreteness, suppose that T is the Wald-like test statistic for the omnibus null hypothesis H_0: $\beta_1 = \cdots = \beta_k = 0$ in a robust regression, calculated using the estimated asymptotic covariance matrix for the regression coefficients. That is, let $\underset{(k\times k)}{\mathbf{V}_{11}}$ contain the rows and columns of the estimated asymptotic covariance matrix $\widehat{\mathcal{V}}(\mathbf{b})$ that pertain to the k slope coefficients $\mathbf{b}_1 = [B_1, \ldots, B_k]'$. We can write the null hypothesis as H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$. Then the test statistic is

$$T = \mathbf{b}_1' \mathbf{V}_{11}^{-1} \mathbf{b}_1$$

[19] The negative correlation of the coefficients reflects the *positive* correlation between income and education (see Section 9.4.4). The hint of bimodality in the distribution of the income coefficient suggests the possible presence of influential observations. See the discussion of Duncan's regression in Section 4.6.

[20] For the use of the bootstrap to estimate bias, see Exercise 21.4.

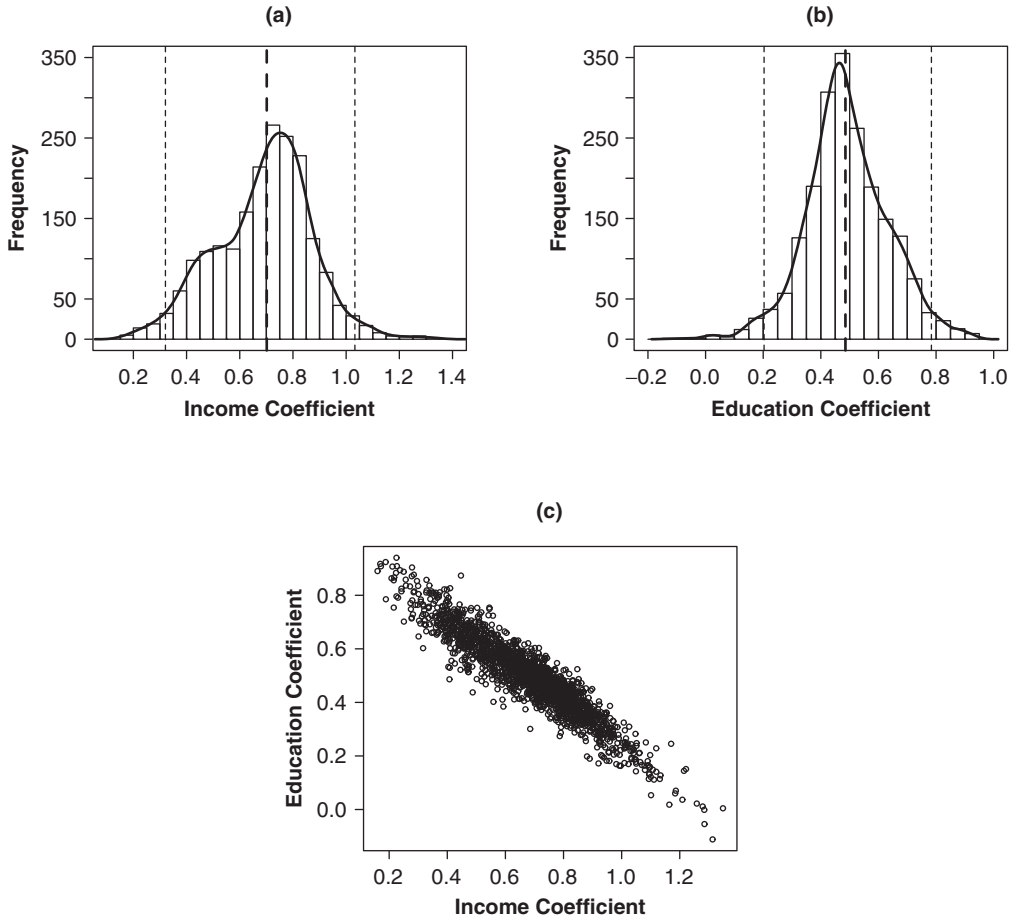

Figure 21.3 Panels (a) and (b) show histograms and kernel density estimates for the $r = 2{,}000$ bootstrap replicates of the income and education coefficients in Duncan's occupational prestige regression. The regression model was fit by *M* estimation using the Huber weight function. Panel (c) shows a scatterplot of the income and education coefficients for the 2,000 bootstrap samples.

We could compare the obtained value of this statistic to the quantiles of χ^2_k, but we are loath to do so because we do not trust the asymptotics. We can, instead, construct the sampling distribution of the test statistic nonparametrically, using the bootstrap.

Let $T^*_b \equiv t(\mathbf{z}^*_b)$ represent the test statistic calculated for the bth bootstrap sample, $\mathbf{z}^*_b$. We have to be careful to draw a proper analogy here: Because the original-sample estimates play the role of the regression parameters in the bootstrap "population" (i.e., the original sample), the bootstrap analog of the null hypothesis—to be used with each bootstrap sample—is H_0: $\beta_1 = B_1, \ldots, \beta_k = B_k$. The bootstrapped test statistic is, therefore,

$$T^*_b = (\mathbf{b}^*_{b1} - \mathbf{b}_1)' \mathbf{V}^{*-1}_{b,11} (\mathbf{b}^*_{b1} - \mathbf{b}_1)$$

Having obtained r bootstrap replications of the test statistic, the bootstrap estimate of the p-value for H_0 is simply[21]

$$\widehat{p}^* = \frac{\#_{b=1}^{r}(T_b^* \geq T)}{r}$$

Note that for this chi-square-like test, the p-value is entirely from the upper tail of the distribution of the bootstrapped test statistics.

> Bootstrap hypothesis tests proceed by constructing an empirical sampling distribution for the test statistic. If T represents the test statistic computed for the original sample, and T_b^* is the test statistic for the bth of r bootstrap samples, then (for a chi-square-like test statistic) the p-value for the test is $\#(T_b^* \geq T)/r$.

21.5 Bootstrapping Complex Sampling Designs

One of the great virtues of the bootstrap is that it can be applied in a natural manner to more complex sampling designs. If, for example, the population is divided into S *strata*, with n_s observations drawn from stratum s, then bootstrap samples can be constructed by resampling n_s observations with replacement from the sth stratum. Likewise, if observations are drawn into the sample in *clusters* rather than individually, then the bootstrap should resample clusters rather than individuals. We can still calculate estimates and test statistics in the usual manner using the bootstrap to assess sampling variation in place of the standard formulas, which are appropriate for independent random samples but not for complex survey samples.

When different observations are selected for the sample with unequal probabilities, it is common to take account of this fact by differentially weighting the observations in inverse proportion to their probability of selection.[22] Thus, for example, in calculating the (weighted) sample mean of a variable Y, we take

$$\overline{Y}^{(w)} = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}$$

and to calculate the (weighted) correlation of X and Y, we take

$$r_{XY}^{(w)} = \frac{\sum w_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{[\sum w_i (X_i - \overline{X})^2][\sum w_i (Y_i - \overline{Y})^2]}}$$

Other statistical formulas can be adjusted analogously.[23]

The case weights are often scaled so that $\sum w_i = n$, but simply incorporating the weights in the usual formulas for standard errors does not produce correct results. Once more, the bootstrap

[21] There is a subtle point here: We use the sample estimate $\mathbf{b}_1$ in place of the hypothesized parameter $\boldsymbol{\beta}_1^{(0)}$ to calculate the bootstrapped test statistic T_b^* *regardless* of the hypothesis that we are testing—because in the central bootstrap analogy $\mathbf{b}_1$ stands in for $\boldsymbol{\beta}_1$ (and the bootstrapped sampling distribution of the test statistic is computed under the assumption that the hypothesis is *true*). See Exercise 21.5 for an application of this test to Duncan's regression.

[22] These "case weights" are to be distinguished from the variance weights used in weighted-least-squares regression (see Section 12.2.2).

[23] See Exercise 21.6.

provides a straightforward solution: Draw bootstrap samples in which the probability of inclusion is proportional to the probability of inclusion in the original sample, and calculate bootstrap replicates of the statistics of interest using the case weights.

The essential "trick" of using the bootstrap in these (and other) instances is to resample from the data in the same way as the original sample was drawn from the population. Statistics are calculated for each bootstrap replication in the same manner as for the original sample.

> The bootstrap can be applied to complex sampling designs (involving, e.g., stratification, clustering, and case-weighting) by resampling from the sample data in the same manner as the original sample was selected from the population.

Social scientists frequently analyze data from complex sampling designs as if they originate from independent random samples (even though there are often non-negligible dependencies among the observations) or employ ad hoc adjustments (e.g., by weighting). A tacit defense of common practice is that to take account of the dependencies in complex sampling designs is too difficult. The bootstrap provides a simple solution.[24]

21.6 Concluding Remarks

If the bootstrap is so simple and of such broad application, why isn't it used more in the social sciences? Beyond the problem of lack of familiarity (which surely can be remedied), there are, I believe, three serious obstacles to increased use of the bootstrap:

1. Common practice—such as relying on asymptotic results in small samples or treating dependent data as if they were independent—usually *understates* sampling variation and makes results look stronger than they really are. Researchers are understandably reluctant to report honest standard errors when the usual calculations indicate greater precision. It is best, however, not to fool yourself, regardless of what you think about fooling others.
2. Although the conceptual basis of the bootstrap is intuitively simple and although the calculations are straightforward, to apply the bootstrap it is necessary to write or find suitable statistical software. There is some bootstrapping software available, but the nature of the bootstrap—which adapts resampling to the data-collection plan and statistics employed in an investigation—apparently precludes full generality and makes it difficult to use traditional statistical computer packages. After all, researchers are not tediously going to draw 2,000 samples from their data unless a computer program can fully automate the process. This impediment is much less acute in programmable statistical computing environments.[25]
3. Even with good software, the bootstrap is computationally intensive. This barrier to bootstrapping is more apparent than real, however. Computational speed is central to the

[24] Alternatively, we can use sampling-variance estimates that are appropriate to complex survey samples—a subject that is beyond the scope of this book. See, for example, Skinner, Holt, and Smith (1989).

[25] See, for example, the bootstrapping software for the S (R and S-PLUS) statistical-computing environment described by Efron and Tibshirani (1993, app.) and by Davison and Hinkley (1997, chap. 11). Bootstrapping facilities are also provided in the Stata programming environment.

exploratory stages of data analysis: When the outcome of one of many small steps immediately affects the next, rapid results are important. This is why a responsive computing environment is especially useful for regression diagnostics, for example. It is not nearly as important to calculate standard errors and p-values quickly. With powerful, yet relatively inexpensive, desktop computers, there is nothing to preclude the machine from cranking away unattended for a few hours (although that is rarely necessary—a few minutes is more typical). The time and effort involved in a bootstrap calculation are usually small compared with the totality of a research investigation—and are a small price to pay for accurate and realistic inference.

Exercises

Exercise 21.1. *Show that the mean of the n^n bootstrap means is the sample mean

$$E^*(\overline{Y}^*) = \frac{\sum_{b=1}^{n^n} \overline{Y}^*_b}{n^n} = \overline{Y}$$

and that the standard deviation (standard error) of the bootstrap means is

$$\mathrm{SE}^*(\overline{Y}^*) = \sqrt{\frac{\sum_{b=1}^{n^n} (\overline{Y}^*_b - \overline{Y})^2}{n^n}} = \frac{S}{\sqrt{n-1}}$$

where $S = \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2/(n-1)}$ is the sample standard deviation. (*Hint*: Exploit the fact that the mean is a linear function of the observations.)

Exercise 21.2. The jackknife: The "jackknife" (suggested for estimation of standard errors by Tukey, 1958) is an alternative to the bootstrap that requires less computation, but that often does not perform as well and is not quite as general. Efron and Tibshirani (1993, chap. 11) show that the jackknife is an approximation to the bootstrap. Here is a brief description of the jackknife for the estimator $\widehat{\theta}$ of a parameter θ:

1. Divide the sample into m independent groups. In most instances (unless the sample size is very large), we take $m = n$, in which case each observation constitutes a "group." If the data originate from a cluster sample, then the observations in a cluster should be kept together.
2. Recalculate the estimator omitting the jth group, $j = 1, \ldots, m$, denoting the resulting value of the estimator as $\widehat{\theta}_{(-j)}$. The *pseudo-value* associated with the jth group is defined as $\widehat{\theta}^*_j \equiv m\widehat{\theta} - (m-1)\widehat{\theta}_{(-j)}$.
3. The average of the pseudo-values, $\widehat{\theta}^* \equiv (\sum_{j=1}^{m} \widehat{\theta}^*_j)/m$ is the jackknifed estimate of θ. A jackknifed $100(1-a)\%$ confidence interval for θ is given by

$$\theta = \widehat{\theta}^* \pm t_{a/2,m-1} \frac{S^*}{\sqrt{n}}$$

where $t_{a/2,m-1}$ is the critical value of t with probability $a/2$ to the right for $m-1$ degrees of freedom, and $S^* \equiv \sqrt{\sum_{j=1}^{m} (\widehat{\theta}^*_j - \widehat{\theta}^*)^2/(m-1)}$ is the standard deviation of the pseudo-values.

(a) *Show that when the jackknife procedure is applied to the mean with $m = n$, the pseudo-values are just the original observations, $\widehat{\theta}^*_i = Y_i$; the jackknifed estimate $\widehat{\theta}^*$ is, therefore,

the sample mean $\overline{Y}$; and the jackknifed confidence interval is the same as the usual t confidence interval.

(b) Demonstrate the results in part (a) numerically for the contrived "data" in Table 21.3. (These results are peculiar to linear statistics like the mean.)

(c) Find jackknifed confidence intervals for the Huber M estimator of Duncan's regression of occupational prestige on income and education. Compare these intervals with the bootstrap and normal-theory intervals given in Table 21.5.

Exercise 21.3. Random versus fixed resampling in regression:

(a) Recall (from Chapter 2), Davis's data on measured and reported weight for 101 women engaged in regular exercise. Bootstrap the least-squares regression of reported weight on measured weight, drawing $r = 1,000$ bootstrap samples using (1) random-X resampling and (2) fixed-X resampling. In each case, plot a histogram (and, if you wish, a density estimate) of the 1,000 bootstrap slopes, and calculate the bootstrap estimate of standard error for the slope. How does the influential outlier in this regression affect random resampling? How does it affect fixed resampling?

(b) Randomly construct a data set of 100 observations according to the regression model $Y_i = 5 + 2x_i + \varepsilon_i$, where $x_i = 1, 2, \ldots, 100$, and the errors are independent (but seriously heteroscedastic), with $\varepsilon_i \sim N(0, x_i^2)$. As in (a), bootstrap the least-squares regression of Y on x, using (1) random resampling and (2) fixed resampling. In each case, plot the bootstrap distribution of the slope coefficient, and calculate the bootstrap estimate of standard error for this coefficient. Compare the results for random and fixed resampling. For a few of the bootstrap samples, plot the least-squares residuals against the fitted values. How do these plots differ for fixed versus random resampling?

(c) Why might random resampling be preferred in these contexts, even if (as is *not* the case for Davis's data) the X values are best conceived as fixed?

Exercise 21.4. Bootstrap estimates of bias: The bootstrap can be used to estimate the bias of an estimator $\widehat{\theta}$ of a parameter θ, simply by comparing the mean of the bootstrap distribution $\overline{\theta}^*$ (which stands in for the expectation of the estimator) with the sample estimate $\widehat{\theta}$ (which stands in for the parameter); that is, $\widehat{\text{bias}} = \overline{\theta}^* - \widehat{\theta}$. (Further discussion and more sophisticated methods are described in Efron and Tibshirani, 1993, chap. 10.) Employ this approach to estimate the bias of the maximum-likelihood estimator of the variance, $\widehat{\sigma}^2 = \sum(Y_i - \overline{Y})^2/n$, for a sample of $n = 10$ observations drawn from the normal distribution $N(0, 100)$. Use $r = 500$ bootstrap replications. How close is the bootstrap bias estimate to the theoretical value $-\sigma^2/n = -100/10 = -10$?

Exercise 21.5. *Test the omnibus null hypothesis H_0: $\beta_1 = \beta_2 = 0$ for the Huber M estimator in Duncan's regression of occupational prestige on income and education.

(a) Base the test on the estimated asymptotic covariance matrix of the coefficients.

(b) Use the bootstrap approach described in Section 21.4.

Exercise 21.6. Case weights:

(a) *Show how case weights can be used to "adjust" the usual formulas for the least-squares coefficients and their covariance matrix. How do these case-weighted formulas compare with those for weighted-least-squares regression (discussed in Section 12.2.2)?

(b) Using data from a sample survey that employed disproportional sampling and for which case weights are supplied, estimate a least-squares regression (1) ignoring the case weights; (2) using the case weights to estimate both the regression coefficients and their standard errors (rescaling the case weights, if necessary, so that they sum to the sample size); and (3) using the case weights but estimating coefficient standard errors with the bootstrap. Compare the estimates and standard errors obtained in (1), (2), and (3).

Exercise 21.7. *Bootstrapping time-series regression: Bootstrapping can be adapted to time series regression but, as in the case of fixed-X resampling, the procedure makes strong use of the model fit to the data—in particular, the manner in which serial dependency in the data is modeled. Suppose that the errors in the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ follow a first-order autoregressive process (see Chapter 16), $\varepsilon_i = \rho\varepsilon_{i-1} + \upsilon_i$; the υ_i are independently and identically distributed with zero expectations and common variance σ_υ^2. Suppose further that we use the method of maximum likelihood to obtain estimates $\widehat{\rho}$ and $\widehat{\boldsymbol{\beta}}$. From the residuals $\mathbf{e} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$, we can estimate υ_i as $V_i = E_i - \widehat{\rho}E_{i-1}$ for $i = 2, \ldots, n$; by convention, we take $V_1 = E_1$. Then, for each bootstrap replication, we sample n values with replacement from the V_i; call them V_{b1}^*, $V_{b2}^*, \ldots, V_{bn}^*$. Using these values, we construct residuals $E_{b1}^* = V_{b1}^*$ and $E_{bi}^* = \widehat{\rho}E_{b,\,i-1}^* + V_{bi}^*$ for $i = 2, \ldots, n$; and from these residuals and the original fitted values $\widehat{Y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$, we construct bootstrapped Y values, $Y_{bi}^* = \widehat{Y}_i + E_{bi}^*$. The Y_{bi}^* are used along with the original $\mathbf{x}_i'$ to obtain bootstrap replicates $\widehat{\boldsymbol{\beta}}_b^*$ of the ML coefficient estimates. (Why are the $\mathbf{x}_i'$ treated as fixed?) Employ this procedure to compute standard errors of the coefficient estimates in the time-series regression for the Canadian women's crime-rate data (discussed in Chapter 16), using an AR(1) process for the errors. Compare the bootstrap standard errors with the usual asymptotic standard errors. Which standard errors do you prefer? Why? Then describe a bootstrap procedure for a time-series regression model with AR(2) errors, and apply this procedure to the Canadian women's crime-rate regression.

Summary

- Bootstrapping is a broadly applicable, nonparametric approach to statistical inference that substitutes intensive computation for more traditional distributional assumptions and asymptotic results. The bootstrap can be used to derive accurate standard errors, confidence intervals, and hypothesis tests for most statistics.
- Bootstrapping uses the sample data to estimate relevant characteristics of the population. The sampling distribution of a statistic is then constructed empirically by resampling from the sample. The resampling procedure is designed to parallel the process by which sample observations were drawn from the population. For example, if the data represent an independent random sample of size n (or a simple random sample of size n from a much larger population), then each bootstrap sample selects n observations with replacement from the original sample. The key bootstrap analogy is the following: *The population is to the sample as the sample is to the bootstrap samples*.
- Having produced r bootstrap replicates $\widehat{\theta}_b^*$ of an estimator $\widehat{\theta}$, the bootstrap standard error is the standard deviation of the bootstrap replicates:

$$\mathrm{SE}^*(\widehat{\theta}^*) = \sqrt{\frac{\sum_{b=1}^{r}(\widehat{\theta}_b^* - \overline{\theta}^*)^2}{r-1}}$$

where $\overline{\theta}^*$ is the mean of the $\widehat{\theta}_b^*$. In large samples, where we can rely on the normality of $\widehat{\theta}$, a 95% confidence interval for θ is given by $\widehat{\theta} \pm 1.96\ \mathrm{SE}^*(\widehat{\theta}^*)$.

- A nonparametric confidence interval for θ can be constructed from the quantiles of the bootstrap sampling distribution of $\widehat{\theta}^*$. The 95% percentile interval is $\widehat{\theta}^*_{(\mathrm{lower})} < \theta < \widehat{\theta}^*_{(\mathrm{upper})}$, where the $\widehat{\theta}^*_{(b)}$ are the r ordered bootstrap replicates; lower $= .025 \times r$ and upper $= .975 \times r$.
- The lower and upper bounds of percentile confidence intervals can be corrected to improve the accuracy of these intervals.
- Regression models can be bootstrapped by (1) treating the regressors as random and selecting bootstrap samples directly from the observations $\mathbf{z}_i' = [Y_i, X_{i1}, \ldots, X_{ik}]$ or (2) treating the regressors as fixed and resampling from the residuals E_i of the fitted regression model. In the latter instance, bootstrap observations are constructed as $Y^*_{bi} = \widehat{Y}_i + E^*_{bi}$, where the $\widehat{Y}_i$ are the fitted values from the original regression, and the E^*_{bi} are the resampled residuals for the bth bootstrap sample. In each bootstrap sample, the Y^*_{bi} are then regressed on the original Xs. A disadvantage of fixed-X resampling is that the procedure implicitly assumes that the regression model fit to the data is correct and that the errors are identically distributed.
- Bootstrap hypothesis tests proceed by constructing an empirical sampling distribution for the test statistic. If T represents the test statistic computed for the original sample and T^*_b is the test statistic for the bth of r bootstrap samples, then (for a chi-square-like test statistic) the p-value for the test is $\#(T^*_b \geq T)/r$.
- The bootstrap can be applied to complex sampling designs (involving, e.g., stratification, clustering, and case weighting) by resampling from the sample data in the same manner as the original sample was selected from the population.

Recommended Reading

Bootstrapping is a rich topic; the presentation in this chapter has stressed computational procedures at the expense of a detailed account of statistical properties and limitations.

- Although Efron and Tibshirani's (1993) book on the bootstrap contains some relatively advanced material, most of the exposition requires only modest statistical background and is eminently readable.
- Davison and Hinkley (1997) is another statistically sophisticated, comprehensive treatment of bootstrapping.
- A briefer source on bootstrapping addressed to social scientists is Stine (1990), which includes a fine discussion of the rationale of bootstrap confidence intervals.
- Young's (1994) paper and the commentary that follows it focus on practical difficulties in applying the bootstrap.