To illustrate bootstrapping, suppose that you have a dataset containing $N$ observations and an estimator that, when applied to the data, produces certain statistics. You draw, with replacement, $N$ observations from the $N$-observation dataset. In this random drawing, some of the original observations will appear once, some more than once, and some not at all. Using the resampled dataset, you apply the estimator and collect the statistics. This process is repeated many times; each time, a new random sample is drawn and the statistics are recalculated.

This process builds a dataset of replicated statistics. From these data, you can calculate the standard error by using the standard formula for the sample standard deviation

$$\widehat{\text{se}} = \left\{ \frac{1}{k-1} \sum (\widehat{\theta}_i - \overline{\theta})^2 \right\}^{1/2}$$

where $\widehat{\theta}_i$ is the statistic calculated using the $i$th bootstrap sample and $k$ is the number of replications. This formula gives an estimate of the standard error of the statistic, according to Hall and Wilson (1991). Although the average, $\overline{\theta}$, of the bootstrapped estimates is used in calculating the standard deviation, it is not used as the estimated value of the statistic itself. Instead, the original observed value of the statistic, $\widehat{\theta}$, is used, meaning the value of the statistic computed using the original $N$ observations.

You might think that $\overline{\theta}$ is a better estimate of the parameter than $\widehat{\theta}$, but it is not. If the statistic is biased, bootstrapping exaggerates the bias. In fact, the bias can be estimated as $\overline{\theta} - \widehat{\theta}$ (Efron 1982, 33). Knowing this, you might be tempted to subtract this estimate of bias from $\widehat{\theta}$ to produce an unbiased statistic. The bootstrap bias estimate has an indeterminate amount of random error, so this unbiased estimator may have greater mean squared error than the biased estimator (Mooney and Duval 1993; Hinkley 1978). Thus $\widehat{\theta}$ is the best point estimate of the statistic.

The logic behind the bootstrap is that all measures of precision come from a statistic's sampling distribution. When the statistic is estimated on a sample of size $N$ from some population, the sampling distribution tells you the relative frequencies of the values of the statistic. The sampling distribution, in turn, is determined by the distribution of the population and the formula used to estimate the statistic.

Sometimes the sampling distribution can be derived analytically. For instance, if the underlying population is distributed normally and you calculate means, the sampling distribution for the mean is also normal but has a smaller variance than that of the population. In other cases, deriving the sampling distribution is difficult, as when means are calculated from nonnormal populations. Sometimes, as in the case of means, it is not too difficult to derive the sampling distribution as the sample size goes to infinity ($N \to \infty$). However, such asymptotic distributions may not perform well when applied to finite samples.

If you knew the population distribution, you could obtain the sampling distribution by simulation: you could draw random samples of size $N$, calculate the statistic, and make a tally. Bootstrapping does precisely this, but it uses the observed distribution of the sample in place of the true population distribution. Thus the bootstrap procedure hinges on the assumption that the observed distribution is a good estimate of the underlying population distribution. In return, the bootstrap produces an estimate, called the bootstrap distribution, of the sampling distribution. From this, you can estimate the standard error of the statistic, produce confidence intervals, etc.

The accuracy with which the bootstrap distribution estimates the sampling distribution depends on the number of observations in the original sample and the number of replications in the bootstrap. A crudely estimated sampling distribution is adequate if you are only going to extract, say, a standard error. A better estimate is needed if you want to use the 2.5th and 97.5th percentiles of the distribution to produce a 95% confidence interval. To extract many features simultaneously about the distribution,

an even better estimate is needed. Generally, replications on the order of 1,000 produce very good estimates, but only 50–200 replications are needed for estimates of standard errors. See Poi (2004) for a method to choose the number of bootstrap replications.

## Regression coefficients

### ▷ Example 1

Let's say that we wish to compute bootstrap estimates for the standard errors of the coefficients from the following regression:

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)
. regress mpg weight gear foreign
```

| Source | SS | df | MS | | |
|---|---|---|---|---|---|
| Model | 1629.67805 | 3 | 543.226016 | Number of obs | = 74 |
| Residual | 813.781411 | 70 | 11.6254487 | F( 3, 70) | = 46.73 |
| | | | | Prob > F | = 0.0000 |
| | | | | R-squared | = 0.6670 |
| | | | | Adj R-squared | = 0.6527 |
| Total | 2443.45946 | 73 | 33.4720474 | Root MSE | = 3.4096 |

| mpg | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| weight | -.006139 | .0007949 | -7.72 | 0.000 | -.0077245 | -.0045536 |
| gear_ratio | 1.457113 | 1.541286 | 0.95 | 0.348 | -1.616884 | 4.53111 |
| foreign | -2.221682 | 1.234961 | -1.80 | 0.076 | -4.684735 | .2413715 |
| _cons | 36.10135 | 6.285984 | 5.74 | 0.000 | 23.56435 | 48.63835 |

To run the bootstrap, we simply prefix the above regression command with the `bootstrap` command (specifying its options before the colon separator). We must set the random-number seed before calling `bootstrap`.

```
. bootstrap, reps(100) seed(1): regress mpg weight gear foreign
(running regress on estimation sample)
Bootstrap replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
```

Linear regression

| | |
|---|---|
| Number of obs | = 74 |
| Replications | = 100 |
| Wald chi2(3) | = 111.96 |
| Prob > chi2 | = 0.0000 |
| R-squared | = 0.6670 |
| Adj R-squared | = 0.6527 |
| Root MSE | = 3.4096 |

| mpg | Observed Coef. | Bootstrap Std. Err. | z | P>\|z\| | Normal-based [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| weight | -.006139 | .0006498 | -9.45 | 0.000 | -.0074127 | -.0048654 |
| gear_ratio | 1.457113 | 1.297786 | 1.12 | 0.262 | -1.086501 | 4.000727 |
| foreign | -2.221682 | 1.162728 | -1.91 | 0.056 | -4.500587 | .0572236 |
| _cons | 36.10135 | 4.71779 | 7.65 | 0.000 | 26.85465 | 45.34805 |