Bootstrapping

1. Introduction

The bootstrap is a data-based simulation based method used for statistical inference. The word *bootstrap* comes from the phrase *to pull oneself up by the bootstraps*, linked to the Adventures of Baron Munchausen by Rudolph Erich Raspe. After falling in a deep lake, the Baron had no means of getting up, except by pulling his own bootstraps Simonoff et al. (1994, p.5).

The usual approach to statistical analysis relies on traditional methods that are strictly valid only if the sample size is infinitely large. The main problem, some measure of accuracy is not available for small samples. Bootstrap methods are a good alternative where analytic distributional approximations are difficult. Bootstrap is a method for estimating the distribution of an estimator by re-sampling the data available.

In bootstrapping we treat the sample as the population. In doing so, we take a large number of *re-samples* from the sample. Even though the re-samples have the same number observations as the original sample, with replacement within each draw, there is a possibility of an observation repeating or not being drawn at all in the next re-sample. Because, the observations of these re-samples vary slightly, a statistic $\hat{\theta}^*$, calculated from the first re-sample, will be slightly different from $\hat{\theta}^*$s calculated using other re-samples. The central idea of bootstrapping is that the distribution calculated using $\hat{\theta}^*$s calculated using re-samples is an estimate of the sampling distribution of $\hat{\theta}$ Mooney and Duval (1993, p.6).

2. Why Bootstrap?

Bootstrap is used in situations when assuming the normality of the errors distribution is not correct or plausible. As an example, in data like the annual Oil production in the U.S States. We won't see a normal distribution as some states like Texas and Alaska will have higher values than others where there might not be any values. Similarly, income per capita, number of conflicts per year are some examples from Mooney and Duval (1993, p.7). In essence, bootstrap allows us to estimate the sampling distribution, without making any assumption about the population distribution, or deriving it explicitly.

3. Bootstrap methods

One problem in applied statistics is the determination of an estimator an a standard error and the determination of confidence intervals (Bergström (n.d.)). However, these kind of problem could be resolved using the re-sampling methods as the bootstrapping.

3.1. **Simple bootstrap.** This considers bootstrap methods that are applicable of data with *i.i.d* values. Let X be a sample with values $x_2, x_2, ..., x_n$ from an unknown population and with probability distribution F. The sample could be used to make inferences about a population characteristic expressed as a parameter θ estimated by a sample $\hat{\theta}$. For each sample x_j we put equal probabilities n^{-1} and we call the empirical distribution denoted $\hat{F}$. The bootstrap can be used to obtain the probability distribution of θ and the variance using $\hat{\theta}$. In this sense, we could recreate the original population with re-sampling with replacement from the sample. The procedure of this method is:

(1) Generate a sample of size n, same size of the original data set with replacement from the empirical distribution.
(2) Compute $\hat{\theta}^*$, the value of θ obtained by using the bootstrap sample in place of the original sample.
(3) Repeat steps 1 and 2 k times.

3.2. **Bootstrap methods for time series.** The problem with time series is the dependence presented in data. In this situation the simple re-sampling procedure will fails. However, there are several extensions of the method that can be done. We could mention a couple pf these, like the block bootstrap, the residual bootstrap and the autoregressive-sieve bootstrap. For example, Block bootstrap will try to recreate pseudo data and replicate the dependence in data by re-sampling blocks of data instead of simple observations. The re-sampling is made with sampling with replacement from the blocks to create time series in which we were able to perform statistics of our interest.

3.3. **Block bootstrap methods.** Common block bootstrap methods are the non-overlapping block bootstrap, the moving block bootstrap, the circular block bootstrap and the stationary bootstrap methods.

(1) NBB. The non-overlapping block bootstrap divides the data set X_n of size n into b non-overlapping blocks of length l, where we suppose $1 < l < n$. We define a collection of non-overlapping blocks $B_1, B_2, ..., B_b$ of length "l" contained in X_n. The NBB samples are then generated by selecting b blocks at random from the collection. Stitching these blocks together in the order they were picked will give the bootstrapped sample the collection
(2) MBB. The moving block bootstrap is the moving block bootstrap method which allows the blocks to overlap. An advantage of this method compared to the NBB method is that by allowing overlapping blocks we have a wider range of blocks to sample from.
(3) CBB. The circular block bootstrap has the main purpose to remove the edge effect of uneven weighting of the observations at the beginning and at the end in the MBB method. The CBB blocks are defined as in the MBB method, i.e. overlapping, but uses an end-to-start wrap around of the data around a circle to make additional blocks
(4) SB. The stationary bootstrap differs from the other three earlier mentioned methods in the sense that block length is not fixed but a random variable geometrically distributed with expected value "l". Because of the random block length, the number of blocks is also random.

3.4. **Choosing block length.** The actual optima block size depends on the sample size and is correlation structure. There are two procedures: by cross validation or plug-in methods. In the cross validation method we could pick a criterion like the mean squared error and minimize the estimated criterion to get an estimate of the optimal block size. The plug-in method involve deriving an expression of the optimal value and to plug in all unknown parameters in the expression.

3.5. **Confidence intervals.** An alternative to generate confidence intervals is to take the quantiles from the sample. For example if we let $\hat{\theta}_{\alpha/2}$ be the $100_{\alpha}/2^{th}$ percentile of the parameter estimate of interest from B re-sampling replications. The percentile interval with coverage $1-\alpha$ is obtained by the percentiles $[\hat{\theta}_{\alpha/2}, \hat{\theta}_{1-\alpha/2}]$. If we choose $\alpha = 0.05$ and $k = 1000$ the 25^{th} and 975^{th} value of the ordered estimates will form a confidence interval for the estimate $\hat{\theta}$ of confidence level 95%.

4. Bootstrapping a Regression Model

Consider a linear regression model

$$Y = X\beta + \epsilon$$

where Y is a (n x 1) vector of values of the dependent variable, X is a matrix of (n x k) independent variables' data, β is a (k x 1) vector of regression coefficients, and ϵ is a (n x 1) errors terms.

Resampling Residuals.

Step 1: Take random data for X and Y from the population and estimate the $\hat{\beta}$s using OLS.

Step 2: Using these $\hat{\beta}$s and the observed Y values, we calculate the residuals,

$$\hat{\epsilon_i} = Y_i - \hat{Y_i}.$$

Step 3: Now, we re-sample these residuals with replacement. And create a bootstrapped vector of the independent variable

$$Y_b^* = \hat{Y} + \hat{\epsilon_b^*}$$

Step 4: These bootstrapped independent variables are then regressed on the previous (fixed), independent variables and a vector of bootstrapped $\hat{\beta_b^*}$ for this particular re-sample is obtained.

$$\hat{Y_b^*} = X\hat{\beta_b^*} + \hat{\epsilon}$$

Step 3 and 4 are repeated B times, to obtain a (B x k) matrix, with each row representing βs of one re-sample.

Resampling Cases.

Resample entire cases of data, that is resample rows in the data matrix, X. B samples of size n are generated, with a regression model estimated for each of the resampled version. We get a (B x k) matrix of bootstrapped regression coefficients.

If your model has a stochastic element or the errors are heteroscedastic, its advisable to resample the data matrix.

References

Bergström, F. (n.d.). Bootstrap Methods in Time Series Analysis Matematiska institutionen.

Mooney, C. Z. and Duval, R. (1993). *Bootstrapping : A Nonparametric Approach to Statistical Inference*, number no. 07-095 in *Sage University Papers Series*, SAGE Publications, Inc, Newbury Park, Calif.
URL: *http://libproxy.uwyo.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN= live*

Simonoff, J., Efron, B., Tibshirani, R. and Hjorth, J. S. U. (1994). An Introduction to the Bootstrap., *Journal of the American Statistical Association* **89**(428): 1559–1560.