An introduction to Bootstrap Methods

Dimitris Karlis

Department of Statistics

Athens University of Economics

Lefkada, April 2004

17th Conference of Greek Statistical Society

http://stat-athens.aueb.gr/~karlis/lefkada/boot.pdf

Outline

Monte Carlo Approximation

Suppose that the cdf F of the population is known. We want to calculate

$$\mu(F) = \int \phi(y) dF(y)$$

We can approximate it by using

$$\hat{\mu}(F) = \frac{1}{M} \sum_{i=1}^{M} \phi(y_i)$$

where y_i, $i =, \ldots, M$ random variables simulated from F (or just a sample from F).

We know that if $M \to \infty$ then $\hat{\mu}(F) \to \mu(F)$.

What about if F in not known?

Motivating Bootstrap

Remedy: Why not use an estimate of F based on the sample $(x_1, \ldots, x_n)$ at hand?

The most well known estimate of F is the empirical distribution function

$$\hat{F}_n(x) = \frac{\#\text{ observations} \leq x}{n}$$

or more formally

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$$

where $I(A)$ is the indicator function and the subscript n reminds us that it is based on sample of size n.

Example: Empirical Distribution Function

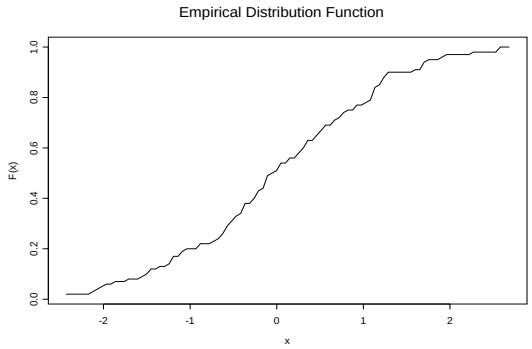

Figure 1: Empirical distribution function from a random sample of size n from a standard normal distribution

Bootstrap Idea

Use as an estimate the quantity $\mu(\hat{F}_n)$ instead of $\mu(F)$. Since $\hat{F}_n$ is a consistent estimate of F (i.e. $\hat{F}_n \to F$ if $n \to \infty$) then $\mu(\hat{F}_n) \to \mu(F)$.

Important: $\mu(\hat{F}_n)$ is an exact result. In practice it is not easy to find it, so we use a Monte Carlo approximation of it.

So, the idea of bootstrap used in practice is the following:

Generate samples from $\hat{F}_n$ and use as an estimate of $\mu(F)$ the quantity the

$$\hat{\mu}(F) = \frac{1}{M}\sum_{i=1}^{M}\phi(y_i^*)$$

where y_i^*, $i =, \ldots, M$ random variables simulated from $\hat{F}_n$.

Simulation from $\hat{F}_n$

Simulating from $\hat{F}_n$ is a relatively easy and straightforward task. The density function $\hat{f}_n$ associated with $\hat{F}_n$ will be the one that gives probability $1/n$ to all observed points x_i, $i = 1, \ldots, n$ and 0 elsewhere.

Note: if some value occurs more than one time then it is given probability larger than $1/n$.

So we sample by selecting randomly with **replacement** observations from the original sample.

A value can occur more than one in the bootstrap sample! Some other values may not occur in the bootstrap sample

A quick view of Bootstrap (1)

- Appeared in 1979 by the seminal paper of Efron. Predecessors existed for a long time
- Popularized in 80's due to the introduction of computers in statistical practice
- It has a strong mathematical background (though not treated here).
- In practice, it is based on simulation but for some few examples there are exact solutions without need of simulation
- While it is a method for improving estimators, it is well known as a method for estimating standard errors, bias and constructing confidence intervals for parameters

A quick view of Bootstrap (2)

- It has minimum assumptions. It is merely based on the assumption that the sample is a good representation of the unknown population
- It is not a black box method. It works for the majority of problems but it may be problematic for some others
- In practice it is computationally demanding, but the progress on computer speed makes it easily available in everyday practice

Types of Bootstrap

- Parametric Bootstrap: We know that F belongs to a parametric family of distributions and we just estimate its parameters from the sample. We generate samples from F using the estimated parameters.
- Non-parametric Bootstrap: We do not know the form of F and we estimate it by $\hat{F}$ the empirical distribution obtained from the data

The general bootstrap algorithm

1. Generate a sample $\mathbf{x}^*$ of size n from $\hat{F}_n$.
2. Compute $\hat{\theta}^*$ for this bootstrap sample
3. Repeat steps 1 and 2, B time.

By this procedure we end up with bootstrap values $\hat{\theta}^* = (\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B)$. We will use these bootstrap values for calculating all the quantities of interest.

Note that $\hat{\theta}^*$ is a sample from the unknown distribution of $\hat{\theta}$ and thus it contains all the information related to $\hat{\theta}$!

An example: Median

Consider data $\mathbf{x} = (x_1, \ldots, x_n)$. We want to find the standard error of the sample median. Asymptotic arguments exist but they refer to huge sample sizes, not applicable in our case if n small. We use bootstrap.

- We generate a sample $\mathbf{x_1}^*$ by sampling with replacement from $\mathbf{x}$. This is our first bootstrap sample
- For this sample we calculate the sample median, denote it as $\hat{\theta}^*_1$
- Repeat steps 1 and 2, B times.

At the end we have B values $\hat{\theta}^* = (\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B)$. This is a random sample from the distribution of the sample median and hence we can use it to approximate every quantity of interest (e.g. mean, standard deviation, percentiles etc). Moreover, a histogram is an estimate of the unknown density of the sample median. We can study skewness etc.

Bootstrap Standard Errors

Denote θ_i^* the bootstrap value from the $i-$th sample, $i = 1, \ldots, B$. The bootstrap estimate of the standard error of $\hat{\theta}$ is calculated as

$$se_B(\hat{\theta}) = \sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_i^* - \hat{\theta}^*\right)^2}$$

where

$$\hat{\theta}^* = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^*$$

This is merely the standard deviation of the bootstrap values.

Bootstrap Estimate of Bias

Similarly an estimate of the bias of $\hat{\theta}$ is obtained as

$$Bias(\hat{\theta}) = \hat{\theta}^* - \hat{\theta}$$

Note that even if $\hat{\theta}$ is an unbiased estimate, since the above is merely an estimate it can be non-zero. So, this estimate must be seen in connection with the standard errors.

Bootstrap Estimate of Covariance

In a similar manner with the standard errors for one parameter we can obtain bootstrap estimates for the covariance of two parameters. Suppose that $\hat{\theta}_1$ and $\hat{\theta}_2$ are two estimates of interest (e.g. in the normal distribution they can be the mean and the variance, in regression setting two of the regression coefficients). Then the bootstrap estimate of covariance is given by

$$Cov_B(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\theta}_{1i}^* - \hat{\theta}_1^*\right)\left(\hat{\theta}_{2i}^* - \hat{\theta}_2^*\right)$$

where $(\hat{\theta}_{1i}^*, \hat{\theta}_{2i}^*)$ are the bootstrap values for the two parameters taken from the i-th bootstrap sample

Example: Covariance between sample mean and variance

Parametric bootstrap. Samples of sizes $n = 20, 200$ were generated from $N(1,1)$ and $Gamma(1,1)$ densities. Both have $\mu = 1$ and $\sigma^2 = 1$. suppose that $\hat{\theta}_1 = \bar{x}$ and $\hat{\theta}_2 = s^2$. Based on $B = 1000$ replications we estimated the covariance between $\hat{\theta}_1$ and $\hat{\theta}_2$.

	Distribution	
	Normal	Gamma
n=20	0.00031	0.0998
n=200	0.0007	0.0104

Table 1: Estimated covariance for sample mean and variance based on parametric bootstrap ($B = 1000$). From theory, for the normal distribution the covariance is 0.

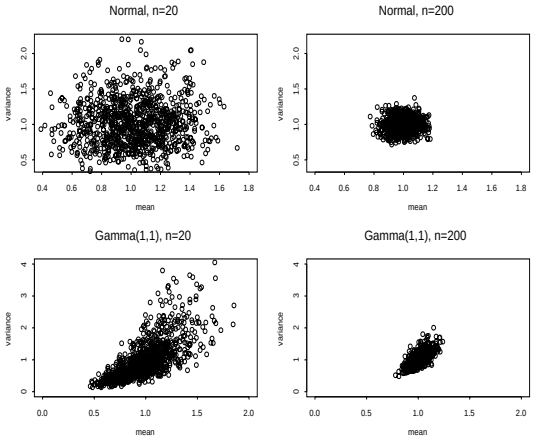

Figure 2: Scatterplot of sample mean and variance based on $B = 1000$ replications.

Simple Bootstrap CI

Use the bootstrap estimate of the standard error and a normality assumption (arbitrary in many circumstances) to construct an interval of the form

$$(\hat{\theta} - Z_{1-a/2} se_B(\hat{\theta}), \hat{\theta} - Z_{1-a/2} se_B(\hat{\theta}))$$

where Z_a denotes the a-th quantile of the standard normal distribution. This is a $(1-a)$ confidence Interval for θ. It implies that we assume that $\hat{\theta}$ follows a Normal distribution.

Percentile CI

The confidence interval is given as

$$\left(\kappa_{a/2}, \kappa_{1-a/2}\right)$$

where κ_a denotes the a-th empirical quantile of the bootstrap values $\hat{\theta}_i^*$.

This is clearly non-symmetric and takes into account the distributional form of estimate.

Percentile-t CI

Improves the simple bootstrap CI in the sense that we do not need the normality assumption. The interval has the form

$$(\hat{\theta} - \zeta_{1-a/2} se_B(\hat{\theta}), \hat{\theta} - \zeta_{a/2} se_B(\hat{\theta}))$$

where ζ_a is the a-th quantile of the values ξ_i, where

$$\xi_i = \frac{\hat{\theta}_i^* - \hat{\theta}}{se(\hat{\theta}_i^*)}$$

Note that for the ξ_i's we need the quantities $se(\hat{\theta}_i^*)$. If we do not know them in closed forms (e.g. asymptotic standard errors) we can use bootstrap to estimate them. The computational burden is double!

Note that ξ_i are studentized values of $\hat{\theta}_i^*$.

Bias Corrected CI

Improves the Percentile bootstrap ci in the sense that we take into account the bias

The bootstrap bias-corrected percentile interval (BC):

$$(\kappa_{p_1}, \kappa_{p_2})$$

where $p_1 = \Phi(z_{\alpha/2} + 2b_0)$ and $p_2 = \Phi(z_{1-\alpha/2} + 2b_0)$ with $\Phi(\cdot)$ the standard normal distribution function,

κ_a is the $a-$quantile of the distribution of the bootstrap values (similar to the notation for the percentile CI)

$$b_0 = \Phi^{-1}\left(\frac{1}{B}\sum_{i=1}^{B} I(\hat{\theta}_i^* \leq \hat{\theta})\right)$$

If the distribution of $\hat{\theta}_i^*$ is symmetric, then $b_0 = 0$ and $p_1 = a/2$ and $p_2 = 1 - a/2$, therefore the simple percentile CI are obtained.

Comparison of CI

Which interval to use? things to take into account:

- Percentile CI are easily applicable in many situations
- Percentile-t intervals need to know $se(\hat{\theta})$.
- In order to estimate consistently extreme percentiles we need to increase B.
- Resulting CI not symmetric (except for the simple CI, which is the less intuitive)
- Note the connection between CI and hypothesis testing!

Example 1: Correlation Coefficient

Consider data (x_i, y_i), $i = 1, \ldots, n$ Pearson's correlation coefficient is given by

$$\hat{\theta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Inference is not so easy, results exist only under the assumption of a bivariate normal population. Bootstrap might be a solution

American Elections

The data refer to $n = 24$ counties in America. They are related to the American presidential elections in 1844. The two variables are: the participation proportion in the election for this county and the difference between the two candidates. The question is whether there is correlation between the two variables. The observed correlation coefficient is $\hat{\theta} = -0.369$

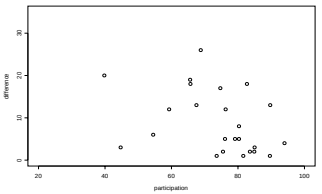

Figure 3: The data for the American Elections

American Elections: Results

From $B = 1000$ replications we found $\hat{\theta}^* = -0.3656$, $se_B(\hat{\theta}) = 0.1801$. The asymptotic standard error given by normal theory as $se_N(\hat{\theta}) = \frac{1-\hat{\theta}^2}{\sqrt{n-3}} = 0.1881$. $Bias(\hat{\theta}) = 0.0042$

simple bootstrap CI	(-0.7228, -0.0167)
Percentile	(-0.6901, 0.0019)
Percentile-t	(-0.6731, 0.1420)
Bias Corrected	(-0.6806, 0.0175)

Table 2: 95% bootstrap confidence intervals for I

More details in the Percentile-t intervals

In this case and since we have an estimate of the standard error of the correlation coefficient (though it is an asymptotic estimate) we can use percentile-t intervals without need to reiterate bootstrap. To do so, from the bootstrap values $\hat{\theta}_i^*$, $i = 1, \ldots, B$, we calculate

$$\xi_i = \frac{\hat{\theta}_i^* - \hat{\theta}}{se(\hat{\theta}_i^*)}$$

where

$$se(\hat{\theta}_i^*) = \frac{1 - \hat{\theta}_i^{*2}}{\sqrt{n-3}}$$

Then we found the quantiles of ξ. Note that the distribution of ξ is skewed, this explains the different left limit of the percentile-t confidence intervals.

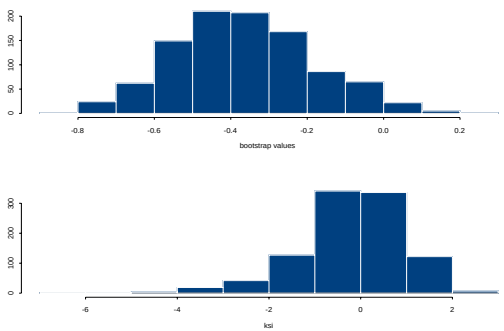

Figure 4: Histogram of the bootstrapped values and the ξ_i's. Skewness is evident in both plots, especially the one for ξ_i's

Example 2: Index of Dispersion

For count data a quantity of interest is $I = s^2/\bar{x}$. For data from a Poisson distribution this quantity is, theoretically, 1. We have accidents for $n = 20$ crossroads in one year period. We want to built confidence intervals for I. The data values are: (1,2,5,0,3,1,0,1,1,2,0,1,8,0,5,0,2,1,2,3).

We use bootstrap by resampling from the observed values. Using $\hat{\theta} = I$ we found ($B = 1000$): $\hat{\theta} = 2.2659$, $\hat{\theta}^* = 2.105$, $\hat{Bias}(\hat{\theta}) = 0.206$, $se_B(\hat{\theta}) = 0.6929$.

simple bootstrap CI	(0.9077, 3.6241)
Percentile	(0.8981, 3.4961)
Percentile-t	(0.8762, 3.4742)
Bias Corrected	(1.0456, 3.7858)

Table 3: 95% bootstrap confidence intervals for I

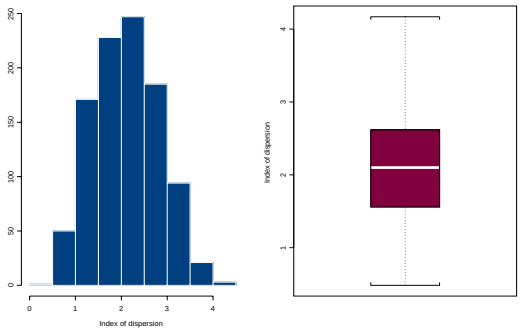

Figure 5: Histogram and boxplot for the index of dispersion. A small skewness is apparent

Some comments

- Asymptotic theory for I is not available; bootstrap is an easy approach to estimate standard errors
- The CI are not the same. This is due to the small skewness of the bootstrap values.
- The sample estimate is biased.
- For the percentile-t intervals we used bootstrap to estimate standard errors for each bootstrap sample.

Hypothesis Tests

Parametric Bootstrap very suitable for hypothesis testing.

Example: We have independent data $(x_1, \ldots, x_n)$, and we know that their distribution is Gamma with some parameters. We wish to test a certain value of the population mean μ: H_0: $\mu = 1$ versus H_1: $\mu \neq 1$.

Standard t-test applicable only via the Central Limit Theorem implying a large sample size. For smaller size the situation is not so easy. The idea is to construct the distribution of the test statistic using bootstrap (parametric).

So, the general algorithm is the following

Hypothesis Tests

- Set the two hypotheses.
- Choose a test statistic T that can discriminate between the two hypotheses. **Important:** We do not care that our statistic has a known distribution under the null hypothesis
- Calculate the observed value t_{obs} of the statistic for the sample
- Generate B samples from the distribution implied by the null hypothesis.
- For each sample calculate the value t_i of the statistic, $i = 1, \ldots, B$.
- Find the proportion of times the sampled values are more extreme than the observed. "Extreme" depends on th form of the alternative.
- Accept or reject according to the significance level.

Hypothesis Tests (2)

More formally, let us assume that we reject the null hypothesis at the right tail of the distribution. Then an approximate p-value is given by

$$\hat{p} = \frac{\sum_{i=1}^{B} I(t_i \geq t_{obs}) + 1}{B+1}$$

$\hat{p}$ is an estimate of the true p-value, we can build confidence intervals for this. A good strategy can be to increase B if $\hat{p}$ is close to the significance level.

Note that some researchers advocate the use of the simpler estimate of the p-value, namely

$$\tilde{p} = \frac{\sum_{i=1}^{B} I(t_i \geq t_{obs})}{B}$$

$\hat{p}$ has better properties than $\tilde{p}$.

Hypothesis Tests (Non-parametric Bootstrap)

The only difficulty is that we do not know the population density. Thus the samples must be taken from $\hat{F}_n$. According to standard hypothesis testing theory, we need the distribution of the test statistic under the null hypothesis. The data are not from the null hypothesis, thus $\hat{F}_n$ is not appropriate. A remedy can be to rescale $\hat{F}_n$ so as to fulfill the null hypothesis. then we take the bootstrap samples from this distribution and we built the distribution of the selected test statistic.

Example

Consider the following data

$$\mathbf{x} = (-0.89, -0.47, 0.05, 0.155, 0.279, 0.775, 1.0016, 1.23, 1.89, 1.96).$$

We want to test $H_0 : \mu = 1$, vs $H_1 : \mu \neq 1$. We select as a test statistic $T = |\bar{x} - 1|$. Several other statistic could be used. Since $\bar{x} = 0.598$ we find $T_{obs} = 0.402$. In order $\hat{F}_{10}$ to represent the null hypothesis we rescale our data so as to have a mean equal to 1. Thus we add in each observation 0.402. The the new sample is $\mathbf{x}_{null} = \mathbf{x} + 0.402$. We resample with replacement from $\hat{F}(\mathbf{x}_{null})$.

Taking $B = 100$ samples we found $\hat{p} = 0.18$.

Important: Rescaling is not obvious in certain hypotheses.

Permutation test

Hypothesis testing via parametric bootstrap is also known as Monte Carlo tests. Alternative testing procedures are the so-called permutation or randomization tests. The idea is applicable when the null hypothesis implies that the data do not have any structure and thus, every permutation of the sample data, under the null hypothesis is equally probable. Then we test whether the observed value is extreme relative to the totality of the permutations. In practice since the permutations are huge we take a random sample from them and we built the distribution of the test statistic under the null hypothesis from them. The main difference from permutation test to bootstrap tests is that in permutation tests we sample without replacement in order to take a permutation of the data.

Permutation test- Example

Consider the data about the American elections in 1844. We want to test the hypothesis that $H_0 : \rho = 0$, vs $H_1 : \rho < 0$. We use as a test statistic r, i.e. the sample correlation coefficient. The observed value is -0.3698. We take $B = 999$ random permutations of the data by fixing the one variable and taking permutations of the other (sampling without replacement). We estimate the p-values as right tail of the distribution. Then an approximate p-value is given by

$$\hat{p} = \frac{\sum_{i=1}^{B} I(t_i \leq t_o bs) + 1}{B+1} = \frac{\sum_{i=1}^{B} I(t_i \leq -0.3698) + 1}{1000} = \frac{39+1}{1000} = 0.04$$

This is an estimate of the true p-value, we can built confidence intervals for this, either via asymptotic results or bootstrap etc. Note that in this case it is not so easy to construct a bootstrap test. Sampling from the null hypothesis is not so easy because we need to transform $\hat{F}_n$ to reflect the null hypothesis! Therefore, permutation tests are complementary to the bootstrap tests.

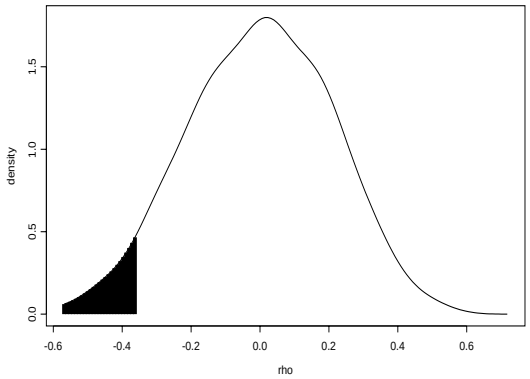

Figure 6: Density estimation for the bootstrapped values. The shadowed area is the area where we reject the null hypothesis

Non-Parametric Bootstrap Hypothesis Tests

Suppose two samples $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_m)$. We wish to test the hypothesis that they have the same mean, i.e. $H_0 : \mu_x = \mu_y$ versus $H_1 : \mu_x \neq \mu_y$. Use as a test statistic $T = |\bar{x} - \bar{y}|$. Under the null hypothesis a good estimate of the population distribution is the combined sample $\mathbf{z} = (x_1, \ldots, x_n, y_1, \ldots, y_m)$. Thus, sample with replacement from $\mathbf{z}$. For each of the B bootstrap samples calculate T_i^*, $i = 1, \ldots B$ Estimate the p-value of the test as

$$\hat{p} = \frac{\sum_{i=1}^{B} I(T_i \geq t_{obs}) + 1}{B+1}$$

Other test statistics are applicable, as for example the well known two-sample t-statistic. A general warning for selecting test statistics is the following: we would like to select a "pivotal" test statistic, i.e. a test statistic for which the distribution does not vary. For example the t-statistic has this property as it is standardized.

Goodness of Fit test using Bootstrap

Parametric Bootstrap suitable for Goodness of fit tests. We wish to test normality, i.e. $H_0 : F = N(\mu, \sigma^2)$ versus $H_1 : F \neq N(\mu, \sigma^2)$. A well known test statistic is the Kolmogorov - Smirnov test statistic $D = max(|\hat{F}_n(x) - F(x)|)$. This has asymptotically and under the null hypothesis a known and tabulated distribution. Bootstrap based test do not use the asymptotic arguments but we take samples from the normal distribution in H_0 (if the parameters are not known, we need to estimate them). For each sample we obtain the value of the test statistic and we construct the distribution of the test statistic. There is no need to use D. Other "distances" can be also use to measure deviations from normality! The test can be used for any distribution!

Failures of Bootstrap

- Small Data sets (because $\hat{F}_n$ is not a good approximation of F)
- Infinite moments (e.g. the mean of the Cauchy distribution).
- Dependence structures (e..g time series, spatial problems). Bootstrap is based on the assumption of independence. Remedies exist
- Estimate extreme values (e.g. 99.99% percentile or $max(X_i)$). The problem is the non-smoothness of the functional under consideration
- Dirty Data: If outliers exist in our sample, clearly we do not sample from a good estimate of F and we add variability in our estimates.
- Unsmooth quantities: There are plenty of theoretical results that relate the success of bootstrap with the smoothness of the functional under consideration
- Multivariate data: When the dimensions of the problem are large, then $\hat{F}_n$ becomes less good as an estimate of F. This may cause problems.

Choice of B

Choice of B depends on

- Computer availability
- Type of the problem: while $B = 1000$ suffices for estimating standard errors, perhaps it is not enough for confidence intervals.
- Complexity of the problem

Variants of Bootstrap

- Smoothed Bootstrap: Instead of using $\hat{f}_n$ we may use a smoothed estimate of it for simulating the bootstrap samples. Such an estimate might be the Kernel Density estimate
- Iterated Bootstrap: For non-smooth functionals (e.g. the median), we perform another bootstrap round using at the bootstrap values $\hat{\theta}_i^*$
- Bayesian Bootstrap: We generate from $\hat{F}$ but the probabilities associated with each observation are not exactly $1/n$ for each bootstrap sample but they vary around this value.

Other Resampling Schemes: The Jackknife

The method was initially introduced as a bias-reduction technique. It is quite useful for estimating standard errors. Let $\hat{\theta}_{(i)}$, denotes the estimate when all the observations except the i-th are used for estimation. Define

$$\hat{\theta}_J = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)}$$

where

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\hat{\theta}_{(i)}$$

$\hat{\theta}_J$ is called the jackknifed version of $\hat{\theta}$ and usually it has less bias than $\hat{\theta}$. It holds that

$$se(\hat{\theta}_J) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}\right)^2}$$

This is a good approximation of $se(\hat{\theta})$ as well.

Other Resampling Schemes: Subsampling

In jackknife we ignore one observation at each time. The samples are created without replacement. Thus, we have $n-1$ different samples. Generalize the idea by ignoring $b \geq 1$ observations at each time. Similar formulas can be derived. Complete enumeration is now difficult: use Monte Carlo methods by taking a sample of them.

This is the idea of subsampling. Subsamples are in fact samples from F and not from $\hat{F}_n$. It can remedy the failure of bootstrap in some cases. Subsamples are of smaller size and thus we need to rescale them (recall the factor $n-1$ in the *se* of the jackknife version.

Split the entire sample in subsamples of equal size.

Bootstrap in Linear Regression

There are two different approaches:

1. Resample with replacement from the observations. Now each observation is the entire vector associated with the original observation. For each
2. Apply bootstrap on the residuals of the model fitted to the original data

The second one is preferable, since the first approach violates the assumption for constant design matrix

Bootstrapping in linear regression removes any distributional assumptions on the residuals and hence allows for inference even if the errors do not follow normal distribution.

Bootstrapping the residuals

Consider the model $Y = \beta X + \epsilon$ using the standard notation.

The bootstrap algorithm is the following

- Fit the model to the original data. Obtain the estimates $\hat{\beta}$ and the residuals form the fitted model $\hat{\epsilon}_i$, $i = 1, \ldots, n$.
- Take a bootstrap sample $\epsilon^* = (\epsilon^*_{\mathbf{1}}, \ldots, \epsilon^*_{\mathbf{n}})$ from the residuals by sampling with replacement.
- Using the design matrix, create the bootstrap values for the response using

$$Y^* = \hat{\beta} X + \epsilon^*$$

- Fit the model using as response Y^* and the design matrix X.
- Keep all the quantities of interest from fitting the model (e.g. MSE, F-statistic, coefficients etc)
- Repeat the procedure B times.

Example

$n = 12$ observations, Y is the wing length of a bird, X is the age of the bird. We want to fit a simple linear regression $Y = \alpha + \beta X$.

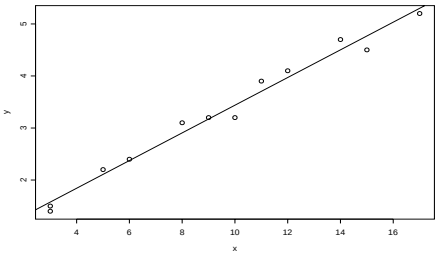

Figure 7: The data and the fitted line

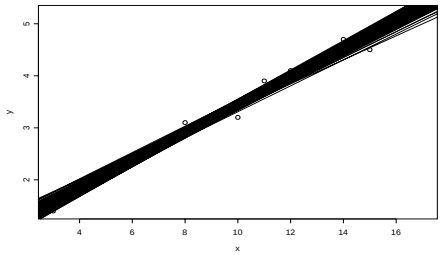

Figure 8: The fitted lines for all the $B = 200$ bootstrap samples

Results

θ	mean	std. err.	95% CI	sample value
$\hat{\alpha}$	0.774	0.106	(0.572 , 0.963)	0.778
$\hat{\beta}$	0.266	0.009	(0.249 , 0.286)	0.266
$\hat{\sigma}$	0.159	0.023	(0.106 , 0.200)	0.178
F -statistic	717.536	288.291	(406.493 , 1566.891)	523.78
R^2	0.984	0.0044	(0.975 , 0.993)	0.9812

Table 4: Bootstrap values for certain quantities. The correlation between $\hat{\alpha}$ and $\hat{\beta}$ was -0.89.

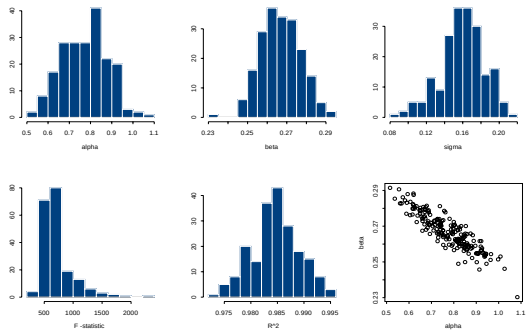

Figure 9: Histogram of the $B = 200$ bootstrap values and a scatterplot for the bootstrap values of $\hat{\alpha}$ and $\hat{\beta}$.

Parametric Bootstrap in Regression

Instead of non-parametric bootstrap we can use parametric bootstrap in a similar fashion. This implies that we assume that the errors follow some distribution (e.g. t or a mixture of normals). Then full inference is available based on bootstrap, while this is very difficult when using classical approaches.

The only change is that the errors for each bootstrap sample are generated from the assumed distribution.

Bootstrap in Principal Components Analysis

Principal Components Analysis is a dimension reduction technique that starts with correlated variables and ends with uncorrelated variables, the principal components, which are in descending order of importance and preserve the total variability. If $\mathbf{X}$ is the vector with the original variables the PC are derived as $\mathbf{Y} = \mathbf{XA}$, where $\mathbf{A}$ is a matrix that contains the normalized eigenvectors from the spectral decomposition of the covariance matrix of $\mathbf{X}$. In order the PCA to me meaningful we need to keep PC less than the original variables. While working with standardized data, a criterion is to select so many components that correspond to eigenvalues greater than 1. A problem of major interest with sample data is how to measure the sampling variability in the eigenvalues of the sample covariance matrix. Bootstrap can be a solution.

Bootstrap in PCA - Example

The data represent the performance of 26 athletes in heptathlon in the Olympic Games of Sidney, 2000. We proceed with PCA based on the correlation matrix. We resample observations with replacement. We used $B = 1000$.

Note: Our approach is non-parametric bootstrap. However we could use parametric bootstrap by assuming a multivariate normal density for the populations and sampling from this multivariate normal model with parameters estimated from the data.

order	mean	st.dev	95% CI	observed value
1	3.0558	0.4403	(2.2548, 3.9391)	2.9601
2	1.6324	0.2041	(1.2331, 2.0196)	1.5199
3	1.0459	0.1830	(0.6833, 1.3818)	1.0505
4	0.6251	0.1477	(0.3733, 0.9416)	0.6464
5	0.3497	0.0938	(0.1884, 0.5441)	0.3860
6	0.1958	0.0658	(0.0846, 0.3401)	0.2778
7	0.0950	0.0413	(0.0306, 0.1867)	0.1592
log-determinant	-4.1194	0.9847	(-6.2859, -2.4292)	-2.9532

Table 5: Bootstrap estimates,standard error and CI, based on $B = 1000$ replications. An idea of the sample variability can be easily deduced.

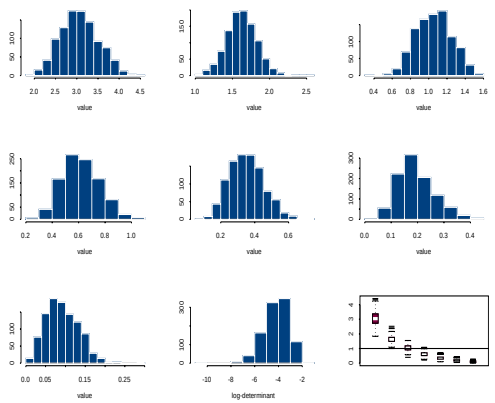

Figure 10: Histogram of the eigenvlaues, the logarithm of the determinant and a boxplot representing all the eigenvalues, based on $B = 1000$ replications

Bootstrap in Kernel Density Estimation (1)

Kernel Density Estimation is a useful tool for estimating probability density functions. Namely we obtain an estimate of the unknown density $f(y)$ as $\hat{f}(y)$ based on a sample $(x_1, \ldots, x_n)$

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - x_i}{h}\right)$$

where $K(\cdot)$ is the kernel function and h is the bandwidth which makes the estimate more smooth ($h \to \infty$) or less smooth ($h \to 0$). There are several choices for the kernel, the standard normal density is the common one. For h we can select it in an optimal way. For certain examples it suffices to use

$$h_{opt} = 1.059 n^{-1/5} min\left(\frac{Q_3 - Q_1}{1.345}, s\right),$$

where Q_1 and Q_3 the first and third quartile and s the sample standard deviation.

We want to create confidence intervals for $\hat{f}(x)$.

Bootstrap in Kernel Density Estimation (2)

We apply bootstrap:

Take B bootstrap samples from the original data. For each sample find $\hat{f}_i(x)$, the subscript denotes the bootstrap sample. Use these values to create confidence intervals for each value of x. This will create a band around $\hat{f}(x)$, which is in fact a componentwise confidence interval and give us information about the curve.

Example: 130 simulated values

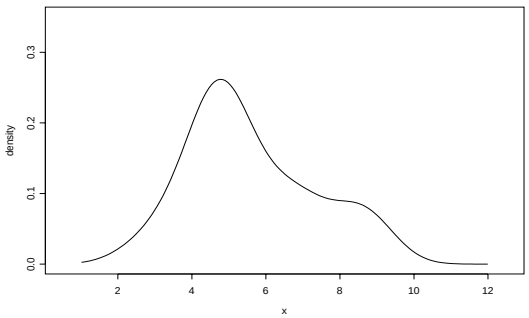

Figure 11: Estimated $\hat{f}(x)$ for the data of our example

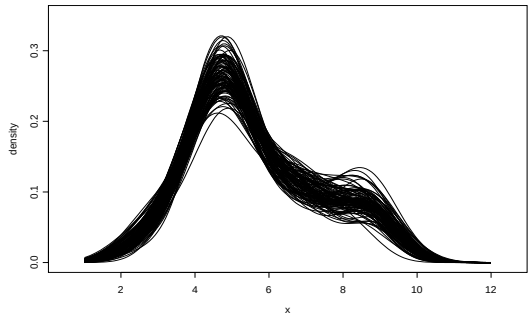

Figure 12: The $B = 100$ bootstrap replicated estimated of $\hat{f}(x)$ for the data of our example

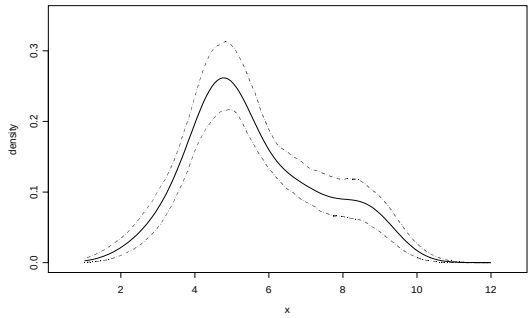

Figure 13: Componentwise 95% confidence intervals based on the percentile bootstrap approach ($B = 100$)

Bootstrap in Time Series

Consider the AR(2) model

$$Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \epsilon_t$$

In the classical case we assume normality for the residuals. If the assumption is not made we can use bootstrap for inference

The algorithm

1. Fit the $AR(2)$ model to the data $(y_1, \ldots, y_T)$. Find $\hat{\beta}_1$, $\hat{\beta}_2$.
2. Obtain the residuals $\hat{\epsilon}_i$ from the fitted model
3. Create the empirical distribution $\hat{F}_T$ of the residuals
4. Start Bootstrap. Simulate a new series by
   - Set $y_i^* = y_i$, $i = 1, 2$, i.e. use as starting points the original observations at that time
   - Construct the entire series using the recursion

     $$y_t^* = \hat{\beta}_1 y_{t-1}^* + \hat{\beta}_2 y_{t-2}^* + \epsilon_t^*,$$

     where ϵ_t^* is drawn from $\hat{F}_T$.
   - Fit the AR(2) model to the bootstrapped series and obtain the parameters
   - Repeat the steps above B times

Example: Annual Peak flows

The data consist of $n = 100$ observations about the annual peak flows in a specific area in the Missouri river in cm^3/sec. The time period examined is 1898-1997.

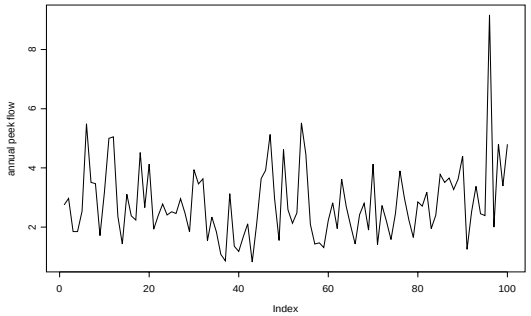

Figure 14: The original series

Example: Annual Peak flows

θ	mean	std. err.	95% CI	sample value
$\hat{\beta}_1$	0.155	0.0964	(-0.032, 0.333)	0.112
$\hat{\beta}_1$	0.042	0.0923	(-0.118, 0.226)	0.063
$\hat{\sigma}^2$	1.669	0.415	(1.087, 2.533)	1.579

Table 6: Bootstrap values for certain quantities. The correlation between $\hat{\alpha}$ and $\hat{\beta}$ was -0.16.

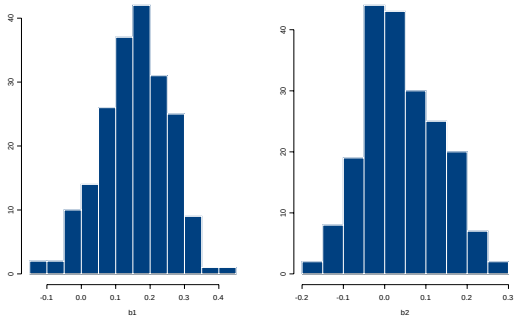

Figure 15: Histogram of the bootstrap values for the parameters ($B = 200$)

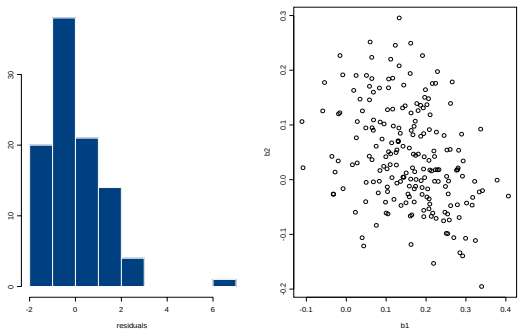

Figure 16: Histogram of the fitted residuals in the sample data and scatterplot of the bootstrap values for the parameters

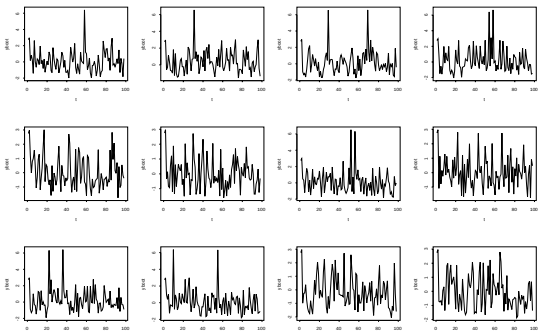

Figure 17: Some plots of the bootstrap series

Bootstrap with dependent Data

Bootstrap is based on independent sampling from $\hat{F}_n$. For dependent data standard bootstrap cannot be applied. The idea is that we need to mimic the dependence of the data.

Remedies

- Moving Block Bootstrap: For example consider the data $(x_1, x_2, \ldots, x_{12})$. We construct 4 blocks of 3 observations in each one, namely $y_1 = (x_1, x_2, x_3)$, $y_2 = (x_4, x_5, x_6)$, $y_3 = (x_7, x_8, x_9)$, and $y_4 = (x_10, x_{11}, x_{12})$. Then we resample blocks y_i instead of the original observation. We keep some part of the dependence, but we loose it when connecting blocks. In some sense we add white noise to the series. Note that dependence for lags ≥ 3 vanishes.
- Overlapping blocks: We built the blocks overlapping, i.e. we define $y_1 = (x_1, x_2, x_3)$, $y_2 = (x_2, x_3, x_4), \ldots, y_{11} = (x_{11}, x_{12}, x_1), y_{12} = (x_{12}, x_1, x_2)$. This adds less white noise but still there is missing dependence

Bootstrap with dependent Data (2)

- Parametric Bootstrap: We apply a parametric model to catch the dependence structure. For example, for linear relationships, by theory, we can find an appropriate AR model that approximates quite well the dependence structure of the series. Or we combine parametric ideas with block bootstrapping, by fitting a model and using block bootstrap to the residuals of that model.
- There are several other most complicated methods suitable for certain types of data (i.e. spatial dependence etc)

More Applications

There are several more applications of bootstrap not treated here. Bootstrap is suitable for a variety of statistical problems like

- Censored Data
- Missing Data problems
- Finite populations
- etc

However before applying bootstrap one must be sure that the bootstrap samples are drawn from a good estimator of the unknown population density. Otherwise bootstrap does not work!

Selected Bibliography

The bibliography on bootstrap increases with high rates during the last years. Some selected textbooks are the following (capturing both the theoretical and the practical aspect of bootstrap)

Efron B., Tibshirani (1993) *An introduction to the Bootstrap.* Marcel and Decker.

Efron, B. (1982) *The jackknife, the bootstrap and other resampling plans.* SIAM. Pensylvania

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statistical Science*, 1, 54–77.

Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Applications.* Cambridge University Press, Cambridge.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap* . Springer

Chernick, M. R. (1999) *Bootstrap Methods: A practitioner's Guide*. Willey

Noreen, E.W. (1989) *Computer Intensive Methods for Testing Hypotheses: An Introduction*. Willey

Hope, A.C.A (1968) A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society*, **B**, 30, 582–598.

Westfall, P.H. and Young, S.S. (1993) *Resampling-Based Multiple Testing*. Wiley

Good, P. (1998). *Resampling Methods: A practical Guide to Data Analysis*. Birkhauser, Boston

THE END