Lecture 3: Bootstrap

Lecturer: Ksenia Kasianova
xeniakasianova@gmail.com

November 20, 2023

Plan

1) What is bootstrap and why we need it?

2) Bootstrap confidence intervals

3) Bias correction

4) Bootstrap hypothesis testing

5) Bootstrap for regression

Principle of obtaining **the sampling distribution**

1. Draw samples from the *population*
2. Compute the statistic of interest for each sample (such as the mean, median, etc.)
3. The distribution of the statistics is the *sampling distribution*

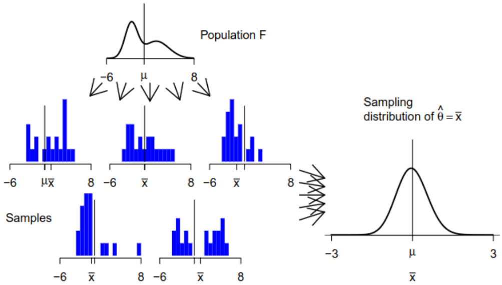


Figure: Sampling distributions are obtained by drawing repeated samples from the population, computing the statistic of interest for each, and collecting (an infinite number of) those statistics as the sampling distribution.

What and why?

If n is large enough for the CLT to have kicked in, we can approximate sampling distributions.

That is, if $X_1, \ldots, X_n$ are a random sample from a distribution "with mean μ and variance σ^2, then

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \text{ is approx } N(0, 1)$$

How big is "large enough"? It depends.

If population dist is close to normal, n can be small and CLT will not have to work hard.

If population distribution is very skewed/asymmetric, n must be much larger to get a normal sampling dist for $\bar{x}$.

Bad idea: sample new data points from the real population.

Cheaper, easier, and faster idea: approximate the sampling distribution with data we already have

Bootstrap principle

1. Draw samples from an estimate of the population.
2. Compute the statistic of interest for each sample.
3. The distribution of the statistics is the bootstrap distribution.

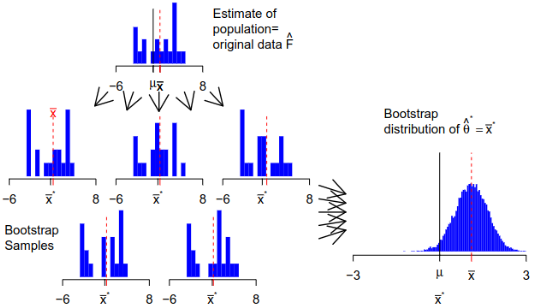


Figure: The bootstrap distribution is obtained by drawing repeated samples from an estimate of the population, computing the statistic of interest for each, and collecting those statistics. The distribution is centered at the observed statistic ($\bar{x}$), not the parameter (μ)

What and why?

What to substitute for F?

- **Nonparametric bootstrap:** The common bootstrap procedure, the non-parametric bootstrap, consists of drawing samples from the empirical distribution $\hat{F}_n$ (with probability $1/n$ on each observation), i.e. drawing samples with replacement from the data.
- **Smoothed Bootstrap:** When we believe the population is continuous, we may draw samples from a smooth population, e.g. from a kernel density estimate of the population.
- **Parametric Bootstrap:** In parametric situations we may estimate parameters of the distribution(s) from the data, then draw samples from the parametric distribution(s) with those parameters.

What and why?

The set-up

1. $x_1, x_2, \ldots, x_n$ is a data sample drawn from a distribution F.
2. u is a statistic computed from the sample.
3. F^* is the empirical distribution of the data (the resampling distribution).
4. $x_1^*, x_2^*, \ldots, x_n^*$ is a resample of the data of the same size as the original sample
5. u^* is the statistic computed from the resample.

Bootstrap principles:

- $F^* \approx F$ and the variation of u is well-approximated by the variation of u^*.
- Sampling with replacement from the data is used: resampling with replacement maintains data structure but reshuffles values, extrapolating to the population

Useful when:

1. data are not normal
2. have unknown statistical properties (e.g., PCA results)
3. lack a standard calculation (e.g., R^2 or coefficient of variation)

(!) Inference, Not Better Estimates

The bootstrap distribution is centered at the observed statistic, not the population parameter, e.g. at $\bar{x}$, not μ.

- We do not use the bootstrap to get better estimates. For example, we cannot use the bootstrap to improve on $\bar{x}$; no matter how many bootstrap samples we take, they are always centered at $\bar{x}$, not μ. We'd just be adding random noise to $\bar{x}$. Instead we use the bootstrap to tell how accurate the original estimate is $=>$ A different approach to estimating the standard error
- We do not use quantiles of the bootstrap distribution of $\hat{\theta}^*$ to estimate quantiles of the sampling distribution of $\hat{\theta}$. Instead, we use the bootstrap distribution to estimate the standard deviation of the sampling distribution, or the expected value of $\hat{\theta} - \theta$.

Estimating the spread and shape of the sampling distribution

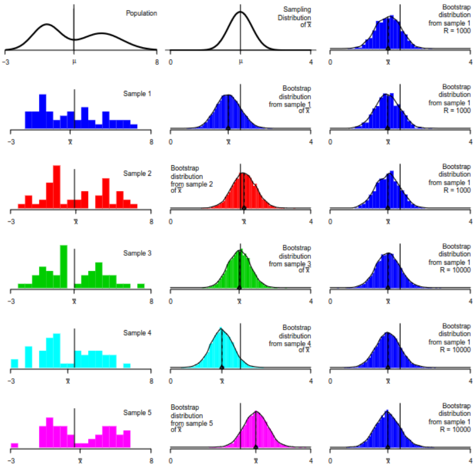


Figure: Bootstrap distribution for the mean, $n = 50$. The left column shows the population and five samples. The middle column shows the sampling distribution for $\bar{X}$, and bootstrap distributions of $\bar{X}^*$ from each sample, with $r = 10^4$. The right column shows more bootstrap distributions from the first sample, three with $r = 1000$ and three with $r = 10^4$.

Small sample limitations – Narrowness bias

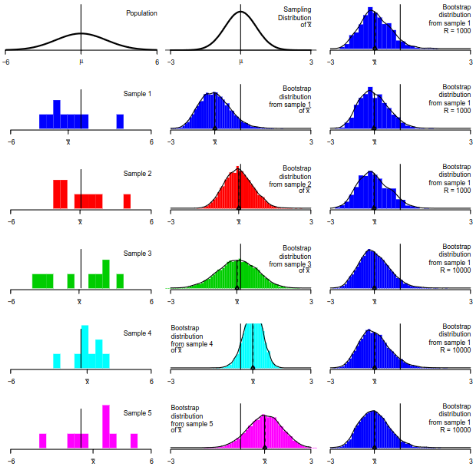


Figure: Bootstrap distributions for the mean, $n = 9$. The left column shows the population and five samples. The middle column shows the sampling distribution for $\bar{X}$, and bootstrap distributions of $\bar{X}^*$ from each sample, with $r = 10^4$. The right column shows more bootstrap distributions from the first sample, three with $r = 1000$ and three with $r = 10^4$.

Summary

The bootstrap distribution reflects the original sample.

Typically for large samples the data represent the population well; for small samples they may not. **Bootstrapping does not overcome the weakness of small samples as a basis for inference.**

Indeed, for the very smallest samples, you may not want to bootstrap; it may be better to make **additional assumptions** such as smoothness or a parametric family.

When there is a lot of data (sampled randomly from a population) we can trust the data to represent the shape and spread of the population; when there is little data we cannot.

Bootstrapping confidence intervals

1. Reverse bootstrap percentile

Let a sample from some distribution with a finite expectation μ be given.

If we knew this distribution we could find $\delta_{.1}$ and $\delta_{.9}$, the 0.1 and 0.9 critical values of δ. Then we'd have

$$P\left(\delta_9 \leq \bar{x} - \mu \leq \delta_{.1} \mid \mu\right) = 0.8 \Leftrightarrow P\left(\bar{x} - \delta_{.9} \geq \mu \geq \bar{x} - \delta_{,1} \mid \mu\right) = 0.8$$

which gives an 80% confidence interval of

$$\left[\bar{x} - \delta_{,1}, \bar{x} - \delta_9\right].$$

As always with confidence intervals, we hasten to point out that the probabilities computed above are probabilities concerning the statistic $\bar{x}$ given that the true mean is μ.

The bootstrap principle offers a practical approach to estimating the distribution of $\delta = \bar{x} - \mu$. It says that we can approximate it by the distribution of

$$\delta^* = \bar{x}^* - \bar{x}$$

where $\bar{x}^*$ is the mean of an empirical bootstrap sample.

$$\left[\bar{x} - \delta_1^*, \bar{x} - \delta_9^*\right]$$

Better to avoid this method since it depends on the bootstrap distribution of $\bar{x}*$ based on a particular sample being a good approximation to the true distribution of $\bar{x}$.

2. Percentile bootstrap

The percentile method directly constructs confidence intervals from the empirical CDF of the bootstrap parameter estimates, $\hat{\theta}_b^*$. The confidence interval is then defined.

$$\left[\hat{\theta}_l^*, \hat{\theta}_u^\star\right]$$

3. t-bootstrap

Let $\widehat{\theta}$ and $s(\widehat{\theta})$ be the estimate of θ and its standard error calculated from the sample. Let $\widehat{\theta}_b^*$ and $s_b(\widehat{\theta})$ be the same quantities from the b th bootstrap sample. Then the b th bootstrap t-statistic is

$$t_{\text{boot },b} = \frac{\widehat{\theta} - \widehat{\theta}_b^*}{s_b(\widehat{\theta})}.$$

As when estimating a population mean, let t_L and t_U be the $\alpha/2$-lower and $\alpha/2$-upper sample quantiles of these t-statistics. Then the confidence interval for θ is

$$\left(\widehat{\theta} + t_L s(\widehat{\theta}), \widehat{\theta} + t_U s(\widehat{\theta})\right)$$

since

$$\begin{aligned} 1 - \alpha &\approx P\left\{t_l \leq \frac{\widehat{\theta} - \widehat{\theta}_b^*}{s_b(\widehat{\theta})} \leq t_U\right\} \\ &\approx P\left\{t_l \leq \frac{\theta - \widehat{\theta}}{s(\widehat{\theta})} \leq t_U\right\} \\ &= P\left\{\hat{\theta} + t_L s(\widehat{\theta}) \leq \theta \leq \hat{\theta} + t_U s(\widehat{\theta})\right\}. \end{aligned}$$

How many observations in bootstrap sample?

Usually n_{boot} is the same as in original sample so that variance of bootstrap sample is equal to the original data

Unless!

We can also modify the sampling to answer *what-if* questions.

Suppose the original sample size was 100, but we draw samples of size 200.

That estimates what would happen with samples of that size—how large standard errors and bias would be, and how wide confidence intervals would be.

Similarly, we can bootstrap with and without stratification and compare the resulting standard errors, to investigate the value of stratification.

How many bootstrap samples?

Question about **Monte Carlo accuracy**!

Increasing the number of re-samples you take, will give you a better estimate of the sampling distribution and more reliable estimate of SE

Rule of thumb: 1000 bootstrap samples for rough approximations, or 10.000 or more for better accuracy

Example 1: In permutation testing we need to estimate the fraction of observations that exceed the observed value => the Monte Carlo standard error is approximately $\sqrt{\hat{p}(1-\hat{p})/r}$, where $\hat{p}$ is the estimated proportion.

Example 2: In bootstrapping, the bias estimate depends on $\overline{\hat{\theta}^*}$, a sample average of r values; the Monte Carlo standard error for this is $s_B/\sqrt{r}$ where s_B is the sample standard deviation of the bootstrap distribution.

Solution: bootstrap the bootstrap!

We can treat the r bootstrap replicates like any old sample, and bootstrap from that sample. For example, to estimate the Monte Carlo SE for the 97.5% quantile of the bootstrap distribution (the endpoint of a bootstrap percentile interval), we draw samples of size r from the r observed values in the bootstrap distribution, compute the quantile for each, and take the standard deviation of those quantiles as the Monte Carlo SE.

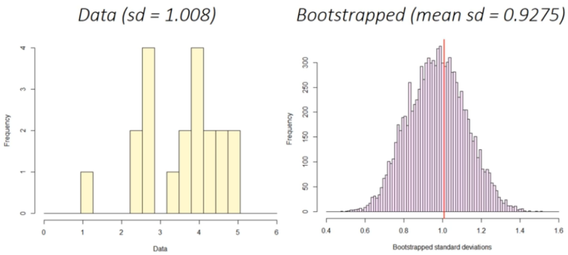
E.g. For two-sided 95% confidence intervals and tests with size 5%. we need $r \geq 15000$ to have accuracy within **10%**

Inherently biased parameters

We can use bootstrap for things that don't even have standard calculation like r-squared or the coefficient of variation

Caution: some parameters are inherently biased during this bootstrapping process

We can sample values multiple times or not at all if we're sampling with replacement but the extreme values tend to be rare compared to values in the center => we tend to underestimate the range of variability in the data and the bootstrapped standard deviations are systematically underestimated



Inherently biased parameters

The bootstrap estimate of bias derives from the plug-in principle. The bias B of a statistic is

$$B = E(\hat{\theta}) - \theta = E_F(\hat{\theta}) - \theta(F)$$

where E_F indicates sampling from F, and $\theta(F)$ is the parameter for population F. The bootstrap substitutes $\hat{F}$ for F, to give

$$\hat{B} = E_{\hat{F}}\left(\hat{\theta}^*\right) - \theta(\hat{F}) = \overline{\hat{\theta}^*} - \hat{\theta}.$$

The bias estimate is the mean of the bootstrap distribution, minus the observed statistic.

Bias correction

We may use the bias estimate to produce a bias-adjusted estimate, $\hat{\theta} - \hat{B} = 2\hat{\theta} - \overline{\hat{\theta}^*}$.

Caution bias estimates can have high variability.

Bias is another reason that we do not use the bootstrap average $\bar{\theta}^*$ in place of $\hat{\theta}$ - it would have double the bias of $\hat{\theta}$.

The bootstrap BCa confidence interval (Efron, 1987) makes use of another kind of bias estimate, the fraction of the bootstrap distribution that is $\leq \hat{\theta}$. It is related to median bias - a statistic is median unbiased if the median of the sampling distribution is θ.

- BCa – the bias-corrected and accelerated bootstrap interval
- Corrects for both bias and skewness
- The bias-correction parameter, z_0, is related to the proportion of bootstrap estimates that are less than the observed statistic.
- The acceleration parameter, *a*, is proportional to the skewness of the bootstrap distribution. You can use the jackknife method to estimate the acceleration parameter.

Bias correction

Simplest: add $\left(sd_{obs} - \overline{sd}_{\text{boot}} \right)$ to each bootstrapped value

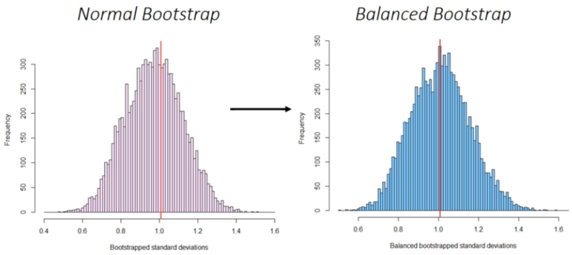


Figure:

We are correcting the offset by adjusting each value so the mean of our bootstrap data is the same as the actual value observed in the original data

Bias correction

We may use the bias estimate to produce a bias-adjusted estimate, $\hat{\theta} - \hat{B} = 2\hat{\theta} - \overline{\hat{\theta}^*}$.

Caution bias estimates can have high variability.

Bias is another reason that we do not use the bootstrap average $\bar{\theta}^*$ in place of $\hat{\theta}$ - it would have double the bias of $\hat{\theta}$.

The bootstrap BCa confidence interval (Efron, 1987) makes use of another kind of bias estimate, the fraction of the bootstrap distribution that is $\leq \hat{\theta}$. It is related to median bias - a statistic is median unbiased if the median of the sampling distribution is θ.

- BCa – the bias-corrected and accelerated bootstrap interval
- Corrects for both bias and skewness
- The bias-correction parameter, z_0, is related to the proportion of bootstrap estimates that are less than the observed statistic.
- The acceleration parameter, a, is proportional to the skewness of the bootstrap distribution. You can use the jackknife method to estimate the acceleration parameter.

Causes of Bias

There are three common causes of bias.

- Bias caused by **nonlinear transformations** for complex statistics – Bias correction would be harmful

Example: estimating the relative risk; $E\left(\hat{p}_1/\hat{p}_2\right) = E\left(\hat{p}_1\right) E\left(1/\hat{p}_2\right) \neq E\left(\hat{p}_1\right)/E\left(\hat{p}_2\right)$.

In this case the median bias is near zero, but the mean bias estimate $\hat{\theta}^* - \hat{\theta}$ can be large and have high variability, and is strongly dependent on how close the denominator is to zero.

Example: s^2 is unbiased but s is not; $E(s) \neq \sqrt{E\left(s^2\right)} = \sigma$.

- Bias due by **optimization**, when one or more parameters are chosen to optimize some measure, then the estimate of that measure is biased. – Bias correction can be helpful.

Example: R^2, sample variance.

- Bias due **lack of model fit** – Bias correction would not be apparent to the bootstrap

Here the bootstrap may not even show that there is bias. It can only quantify the performance of the procedure you actually used, not what you should have used.

Skewness

When the data are skewed, a correct interval is even more asymmetrical than the bootstrap percentile interval—reaching farther toward the long tail

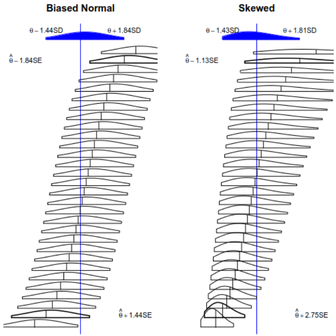


Figure: *Confidence intervals for bias, and acceleration*. The vertical lines correspond to true values of the parameter. On the left the sampling distribution, and bootstrap distributions, are normal with bias (correct interval $(\hat{\theta} - 1.84\mathrm{SE}, \hat{\theta} + 1.44\mathrm{SE})$). The bootstrap percentile interval is asymmetrical in the wrong direction: $(\hat{\theta} - 1.44\mathrm{SE}, \hat{\theta} + 1.84\mathrm{SE})$.
On the right the sampling distribution, and bootstrap distributions, are unbiased with skewness $2/3$ (correct interval should be $(\hat{\theta} - 1.13\mathrm{SE}, \hat{\theta} + 2.75\mathrm{SE})$). The bootstrap percentile interval $(\hat{\theta} - 1.43\mathrm{SE}, \hat{\theta} + 1.81\mathrm{SE})$ is not asymmetrical enough.
A t interval $(\hat{\theta} - 1.64\mathrm{SE}, \hat{\theta} + 1.64\mathrm{SE})$ is even worse.

Hypothesis testing

A bootstrap hypothesis test starts with a test statistic - $t(x)$ (not necessary an estimate of a parameter).

We seek an achieved significance level

$$ASL = \text{Prob}_{H_0}\left\{t\left(\boldsymbol{x}^*\right) \geq t(\boldsymbol{x})\right\}$$

Where the random variable $\boldsymbol{x}^*$ has a distribution specified by the null hypothesis H_0 - denote as F_0. Bootstrap hypothesis testing uses a "plug-in" style to estimate F_0.

Bootstrap Hypothesis Testing $F = G$

- Denote the combined sample by $\boldsymbol{x}$, and its empirical distribution by $\hat{F}_0$.

- Under H_0, $\hat{F}_0$ provides a non parametric estimate for the common population that gave rise to both $\boldsymbol{z}$ and $\boldsymbol{y}$.

1. Draw $\boldsymbol{B}$ samples of size $n+m$ with replacement from $\boldsymbol{x}$. Call the first n observations $\mathbf{z}^*$ and the remaining $m - \boldsymbol{y}^*$

2. Evaluate $t(\cdot)$ on each sample $-t\left(\boldsymbol{x}^{*\boldsymbol{b}}\right)$

3. Approximate ASL_{boot} by

$$\widehat{ASL}_{\text{boot}} = \#\left\{t\left(\boldsymbol{x}^{*\boldsymbol{b}}\right) \geq t(\boldsymbol{x})\right\}/B$$

* In the case that large values of $t\left(\boldsymbol{x}^{*\boldsymbol{b}}\right)$ are evidence against H_0

Testing Equality of Means

- Instead of testing $H_0 : F = G$, we wish to test $\mathrm{H}_0 : \mu_z = \mu_y$, without assuming equal variances. We need estimates of F and G that use only the assumption of common mean

1. Define points $\tilde{z}_i = z_i - \bar{z} + \bar{x}, i = 1, \ldots, n$, and $\tilde{y}_i = y_i - \bar{y} + \bar{x}, i = 1, \ldots, m$. The empirical distributions of $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{y}}$ shares a common mean.

2. Draw $\boldsymbol{B}$ bootstrap samples with replacement $(\boldsymbol{z}^*, \boldsymbol{y}^*)$ from $\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n$ and $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_m$ respectivly

3. Evaluate $t(\cdot)$ on each sample -

$$t\left(\boldsymbol{x}^{*\boldsymbol{b}}\right) = \frac{\bar{z}^* - \bar{y}^*}{\sqrt{\bar{\sigma}_z^* 1/n + \bar{\sigma}_y^* 1/m}}$$

4. Approximate ASL_{boot} by

$$\widehat{ASL}_{\text{boot}} = \#\left\{t\left(\boldsymbol{x}^{*\boldsymbol{b}}\right) \geq t(\boldsymbol{x})\right\}/B$$

Permutation Test VS Bootstrap Hypothesis Testing

- Accuracy: In the two-sample problem, ASL_{perm} is the exact probability of obtaining a test statistic as extreme as the one observed. In contrast, the bootstrap explicitly samples from estimated probability mechanism. $\widehat{ASL}_{\text{boot}}$ has no interpretation as an exact probability.

- Flexibility: When special symmetry isn't required, the bootstrap testing can be applied much more generally than the permutation test. (Like in the two sample problem - permutation test is limited to $H_0 : F = G$, or in the one-sample problem)

Bootstrap for regression

1. Jackknife method

(not actually a bootstrap method)

We want to create "new" data sets from the data we have and use them to compute many values

Then, we can simply compute the variance of the many values §

For jackknife, we create new data sets by:

omitting each sample from one new set

– This means we get N new sets if we have N samples

– Each new set has $N-1$ samples

You could choose a smaller number if N were prohibitively large

2. Pair bootstrap

Idea: resample observations (x_i, y_i) with replacement and estimate

3. Wild Bootstrap Techniques

In regression analysis, the most popular and widely used bootstrap technique is the fixed- x resampling or bootstrapping the residuals. This bootstrapping procedure is based on the ordinary least squares (OLS) residuals summarized as follows.

Step 1. Fit a model $y_i = f(x_i, \beta_{\text{ols}})$ by the OLS method to the original sample of observations to get $\widehat{\beta}_{\text{ols}}$ and hence the fitted model is $\widehat{y}_i = f\left(x_i, \widehat{\beta}_{\text{ols}}\right)$.

Step 2. Compute the OLS residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$ and each residual $\hat{\varepsilon}_i$ has equal probability, $1/n$.

Step 3. Draw a random sample $\varepsilon_1^*, \varepsilon_2^*, \ldots, \varepsilon_n^*$ from $\hat{\varepsilon}_i$ with simple random sampling with replacement and attached to $\hat{y}_i$ for obtaining fixed- x bootstrap values y_i^{*b} where $y_i^{*b} = f\left(x_i, \widehat{\beta}_{\text{ols}}\right) + \varepsilon_i^{*b}$.

Step 4. Fit the OLS to the bootstrapped values y_i^{*b} on the fixed- x to obtain $\hat{\beta}_{\text{ols}*}^{*b}$

Step 5. Repeat Steps 3 and 4 for B times to get $\widehat{\beta}_{\text{ols}}^{*b1}, \ldots, \widehat{\beta}_{\text{ols}}^{*bB}$ where B is the bootstrap replications.

We call this bootstrap scheme $Boot_{\text{ols}}$ since it is based on the OLS method.

Bootstrap for regression

When *heteroscedasticity* is present in the data, the variances of the data are different and neither of these bootstrap schemes can yield efficient estimates of the parameters.

We can modify Step 3 of the OLS bootstrap and kept the other steps unchanged.

For each i, draw a value t_i^*, with replacement, from a distribution with zero mean and unit variance and attached to $\hat{y}_i$ for obtaining fixed- x bootstrap values y_i^{*b}, where $y_i^{*b} = f\left(x_i, \hat{\beta}_{\text{ols}}\right) + t_i^* \hat{\varepsilon}_i / \sqrt{1 - h_{ii}}$ and $h_{\text{ii}} = x_i^T \left(X^T X\right) x_i$ is the i th leverage. Note that the variance of $t_i^* \hat{\varepsilon}_i$ is not constant when the original errors are not homoscedastic.

Therefore, this bootstrap scheme takes into consideration the nonconstancy of the error variances. As an alternative, t_i^* can be chosen, with replacement, from $a_1, a_2, \ldots, a_n$, where

$$a_i = \frac{\hat{\varepsilon}_i - \overline{\hat{\varepsilon}}_i}{\sqrt{n^{-1} \sum_{i=1}^{n} \left(\hat{\varepsilon}_i - \overline{\hat{\varepsilon}}\right)^2}}$$

with $\overline{\hat{\varepsilon}} = n^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_i$.

For a regression model with intercept term, $\overline{\hat{\varepsilon}}_i$ approximately equals zero. This is nonparametric implementation of Wu's bootstrap since the resampling is done from the empirical distribution function of the (normalized) residuals. We call this method Wu's bootstrap and denote it by Boot $_{wu}$.

Summary

– Just another method for statistical inference

– We do not use the bootstrap to get better estimates(!)

Instead we use the bootstrap to tell how accurate the original estimate is

– Can be used for complex statistics

– Non-parametric bootstrap does not help for small samples

– Beware of possible biases

– Works better for skewed data than t interval but still under estimates the skewness