December 25, 2021

Resampling Methods

Motivation

We have so many estimators with the property

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \to_d N\left(0, \sigma^2\right).$$

We can also write $\widehat{\theta}_n \overset{a}{\sim} N\left(\theta, \sigma^2/n\right)$, where $\overset{a}{\sim}$ means "approximately distributed as". Once we have a consistent estimator $\widehat{\sigma}_n$ of σ, the standard error is defined to be $SE = \widehat{\sigma}/\sqrt{n}$. A confidence interval with approximate 95% coverage probability is $\left[\widehat{\theta}_n \pm 1.96 \times SE\right]$. Our strategy for estimating σ^2 was based on the analogue/plug-in principle, i.e., replace population moments/unknown quantities by their sample moments/estimates. We need knowledge of the expression (formula) of σ^2. There are two computation-intensive "resampling" approaches that do the estimation without requiring knowledge of the expression of σ^2.

Suppose we have some (testing) statistic W_n and we need to know its distribution (under the null hypothesis) and calculate its quantile. The approach we took was to find the asymptotic distribution of W_n, which was always standard normal or χ^2. The quantile of the asymptotic distribution can be found easily since it does not depend on any unknown quantity/parameter. We use it as approximation to the true quantile of W_n. Later we will see that there is another approach to approximating the true distribution of W_n.

Resampling methods are now core to modern econometrics. There are least three motivations behind the popularity of the resampling methods.

- Standard errors are hard to get. Suppose $X_1, ..., X_n$ is an i.i.d. random sample with mean μ and variance σ^2. Then the standard error of the sample mean $\widehat{\mu}_n = n^{-1}\sum_{i=1}^n X_i$ is $SE = \widehat{\sigma}_n/\sqrt{n}$ where $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2$. Suppose that X_i is continuous with density f_X. Assume for simplicity that its CDF F_X is strictly increasing. The (population) median is $m = F_X^{-1}(1/2)$, i.e., $\Pr(X_i \leq m) = 1/2$. We order the data: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Define the sample median:

$$\widehat{m}_n = \text{median}\{X_1, ..., X_n\} = \begin{cases} \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2} & \text{if } n \text{ is even} \\ X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd.} \end{cases}$$

  It is known that $\sqrt{n}(\widehat{m}_n - m) \to_d N\left(0, \left(4f_X(m)^2\right)^{-1}\right)$. Constructing a "plug-in" estimator of the asymptotic variance $\left(4f_X(m)^2\right)^{-1}$ requires knowledge of nonparametric econometrics since we need to estimate the density function f_X at a point m. There is also some subtle technical issue with this approach. For this problem, resampling methods come to rescue.

- (Almost) nothing else we can do. Suppose $X_1, ..., X_n$ is an i.i.d. random sample. We want to test $H_0 : X$ is normally distributed, i.e., for some μ and σ, $X_i \sim N\left(\mu, \sigma^2\right)$. Remember that empirical

distribution function $\widehat{F}_n(x) = n^{-1}\sum_{i=1}^n 1(X_i \leq x)$ is consistent for F_X. Indeed, we have a much stronger result: $\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x) - F_X(x)\right| \to_p 0$ (Glivenko-Cantelli theorem). Let $\Phi_{\mu,\sigma}$ be the CDF of $N(\mu,\sigma^2)$. The Kolmogorov–Smirnov test uses the statistic $KS = \sup_{x\in\mathbb{R}}\sqrt{n}\left|\widehat{F}_n(x) - \Phi_{\widehat{\mu},\widehat{\sigma}}(x)\right|$, where $\widehat{\mu} = n^{-1}\sum_{i=1}^n X_i$ and $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \widehat{\mu})^2$. If H_0 is true, both $\widehat{F}_n$ and $\Phi_{\widehat{\mu},\widehat{\sigma}}$ are consistent for F_X and the statistic KS should be small. So a large observed KS is regarded as evidence against H_0. We reject H_0 if $KS > c$. We know that $KS \to_d B$, for some random variable B with a very complicated distribution that depends on unknown parameters. So it is not practically possible to choose c such that $\Pr(B \leq c) = 1-\alpha$. Again for this problem, resampling methods come to rescue.

- For the "traditional" confidence interval $\widehat{\theta}_n \pm 1.96 \times SE$, we know that $\Pr\left(\theta \in \left[\widehat{\theta}_n \pm 1.96 \times SE\right]\right) \to 95\%$ as $n \to \infty$. Actually in many cases we can show that $\Pr\left(\theta \in \left[\widehat{\theta}_n \pm 1.96 \times SE\right]\right) = 95\% + O\left(n^{-1}\right)$, i.e., the error $\Pr\left(\theta \in \left[\widehat{\theta}_n \pm 1.96 \times SE\right]\right) - 95\%$ goes to zero at the rate n^{-1}. Some resampling-based confidence interval $\left[\widehat{\theta}_n + t^*_{2.5\%} \times SE, \widehat{\theta}_n + t^*_{97.5\%} \times SE\right]$ with some "new" critical values $t^*_{2.5\%}$ and $t^*_{97.5\%}$ has the property

$$\Pr\left(\theta \in \left[\widehat{\theta}_n + t^*_{2.5\%} \times SE, \widehat{\theta}_n + t^*_{97.5\%} \times SE\right]\right) = 95\% + O\left(n^{-3/2}\right).$$

  So the error is smaller and the coverage accuracy of the resampling-based confidence interval is much better.

Jackknife

Probably "jackknife" is the first-generation resampling method. Suppose $X_1, ..., X_n$ is an i.i.d. random sample. For simplicity, assume X_i is scalar. An estimator $\widehat{\theta}_n$ can be written as $\widehat{\theta}_n = \varphi_n(X_1, ..., X_n)$, e.g., $\varphi_n(z_1, ..., z_n) = \frac{1}{n}\sum_{i=1}^n z_i$. Suppose we know $\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \to_d N\left(0, \sigma^2\right)$ and we want to estimate σ^2. Now denote $\widehat{\theta}_n^{-j} = \varphi_{n-1}(X_1, ..., X_{j-1}, X_{j+1}, ..., X_n)$, i.e., $\widehat{\theta}_n^{-j}$ is an estimator obtained by removing the j-th observation from the entire sample.

The variation in $\left\{\widehat{\theta}_n^{-j} : j = 1, ..., n\right\}$ should be informative about the population variance of $\widehat{\theta}_n$. Actually it is informative about the population variance of $\widehat{\theta}_{n-1}$. Note that $\widehat{\theta}_{n-1} \overset{a}{\sim} N\left(\theta, \sigma^2/(n-1)\right)$. Denote $\overline{\theta}_n = n^{-1}\sum_{j=1}^n \widehat{\theta}_n^{-j}$. Now it seems reasonable to think of $n^{-1}\sum_{j=1}^n \left(\widehat{\theta}_n^{-j} - \overline{\theta}_n\right)^2$ as an estimate of $\sigma^2/(n-1)$ and $(n-1) \times n^{-1}\sum_{j=1}^n \left(\widehat{\theta}_n^{-j} - \overline{\theta}_n\right)^2$ as an estimate of σ^2. Indeed in many cases one can show

$$(n-1)\sum_{j=1}^n \left(\widehat{\theta}_n^{-j} - \overline{\theta}_n\right)^2 \to_p \sigma^2. \tag{1}$$

The Jackknife standard error is

$$SE = \sqrt{\frac{(n-1)\sum_{j=1}^n \left(\widehat{\theta}_n^{-j} - \overline{\theta}_n\right)^2}{n}}.$$

A jackknife 95% confidence interval is $\left[\widehat{\theta}_n \pm 1.96 \times SE\right]$. If (1) is true, we say that jackknife is

consistent.

Consider the following simple example: for i.i.d. random sample $X_1, ..., X_n$, we use the sample average $\overline{X}_n$ as an estimator of $\mu = \mathbb{E}X_1$. It is known that $\sqrt{n}\left(\overline{X}_n - \mu\right) \to_d N\left(0, \sigma^2\right)$, where $\sigma^2 = \text{Var}\left(X_1\right)$. For this case,

$$\widehat{\theta}_n^{-j} = \frac{1}{n-1}\left(n\overline{X}_n - X_j\right),$$
$$\frac{1}{n}\sum_{j=1}^{n}\widehat{\theta}_n^{-j} = \frac{1}{n\left(n-1\right)}\sum_{j=1}^{n}\left(n\overline{X}_n - X_j\right) = \overline{X}_n,$$

and

$$\widehat{\theta}_n^{-j} - \overline{\theta}_n = \frac{1}{n-1}\left(n\overline{X}_n - X_j\right) - \overline{X}_n = \frac{1}{n-1}\left(\overline{X}_n - X_j\right).$$

We have

$$\left(n-1\right)\sum_{j=1}^{n}\left(\hat{\theta}_{-j} - \overline{\hat{\theta}}\right)^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}_n\right)^2,$$

which is the sample variance that is a consistent and unbiased estimator for σ^2.

Note that unlike the "plug-in" approach, the jackknife approach does not even require knowledge of the expression of σ^2. The limitation of jackknife is that (1) is not always true. For the case of median, (1) fails and jackknife is inconsistent.

Bootstrap

The second-generation resampling method is the bootstrap. First, let us see how bootstrap gets the standard error for estimating the population median and constructs the confidence interval. for i.i.d. random sample $X_1, ..., X_n$, let $\widehat{m}_n = \text{median}\left\{X_1, ..., X_n\right\}$. First we independently draw n observations with replacement from $X_1, ..., X_n$ and get a set of new observations $X_1^{*(1)}, ..., X_n^{*(1)}$. The computer can handle this for us. We repeat this resampling procedure again and again, B times. B is a very large integer. Ideally how B is depends solely on how powerful our computer is. What we have is B bootstrap samples

$$\begin{array}{ccccc}
X_1^{*(1)} & \cdots & X_n^{*(1)} & \Longrightarrow & \widehat{m}_n^{*(1)} = \text{median}\left\{X_1^{*(1)}, ..., X_n^{*(1)}\right\} \\
X_1^{*(2)} & \cdots & X_n^{*(2)} & \Longrightarrow & \widehat{m}_n^{*(2)} = \text{median}\left\{X_1^{*(2)}, ..., X_n^{*(2)}\right\} \\
\vdots & \ddots & \vdots & \vdots & \vdots \\
X_1^{*(B)} & \cdots & X_n^{*(B)} & \Longrightarrow & \widehat{m}_n^{*(B)} = \text{median}\left\{X_1^{*(B)}, ..., X_n^{*(B)}\right\}
\end{array}$$

and for each bootstrap sample, we calculate its sample median.

We use the sample variance of $\widehat{m}_n^{*(1)}, \widehat{m}_n^{*(2)}, ..., \widehat{m}_n^{*(B)}$ as an estimate of the true variance of $\widehat{m}_n$:

$$\widehat{\text{Var}}_{BS}\left(\widehat{m}_n\right) = \frac{1}{B}\sum_{b=1}^{B}\left\{\widehat{m}_n^{*(b)} - \frac{1}{B}\sum_{b=1}^{B}\widehat{m}_n^{*(b)}\right\}^2.$$

Then the bootstrap standard error is $SE = \sqrt{\widehat{\text{Var}}_{BS}\left(\widehat{m}_n\right)}$ and an approximate 95% confidence interval using the bootstrap standard error is $[\widehat{m}_n \pm 1.96 \times SE]$. In fact, there is another seemingly simpler

way to construct the confidence interval. We order the bootstrap sample medians: $\widehat{m}^*_{(1)} \leq \widehat{m}^*_{(2)} \leq \cdots \leq \widehat{m}^*_{(B)}$. Suppose for simplicity $B \times 2.5\%$ and $B \times 97.5\%$ are both integers. A bootstrap percentile confidence interval is simply $\left[\widehat{m}^*_{(B\times 2.5\%)}, \widehat{m}^*_{(B\times 97.5\%)}\right]$.

The bootstrap procedure we just described is called nonparametric bootstrap or empirical bootstrap invented by Professor Bradley Efron in 1979. The (nonparametric) bootstrap takes the sample as the population. A bootstrap sample is obtained by independently drawing n observations from the observed sample with replacement. The bootstrap sample has the same number of observations as the original sample, however some observations appear several times and others never.

Now we summarize the two procedures we introduced. Suppose we have an estimator which is asymptotically normal: $\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \to_d N\left(0, \sigma^2\right)$.

- Bootstrap standard errors
  - Step 1: Draw B independent bootstrap samples. B can be as large as possible. We can take $B = 1000$.
  - Step 2: Estimate θ with each of the bootstrap samples, $\widehat{\theta}^{*(b)}_n$ for $b = 1, ..., B$.
  - Step 3: Estimate the standard error by
$$SE = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}^{*(b)}_n - \overline{\theta}^*\right)^2}$$
where $\overline{\theta}^* = B^{-1}\sum_{b=1}^{B}\widehat{\theta}^{*(b)}_n$.
  - Step 4: The bootstrap standard errors can be used to construct approximate confidence intervals, e.g., if the coverage probability is 95%, a 95% confidence interval is $\left[\widehat{\theta}_n \pm 1.96 \times SE\right]$.
- Bootstrap percentile
  - Step 1: Draw B independent bootstrap samples. B can be as large as possible. We can take $B = 1000$.
  - Step 2: Estimate θ with each of the bootstrap samples, $\widehat{\theta}^{*(b)}_n$ for $b = 1, ..., B$.
  - Step 3: Order the bootstrap replications such that
$$\widehat{\theta}^*_{(1)} \leq \cdots \leq \widehat{\theta}^*_{(B)}.$$
  - Step 4: The lower and upper confidence bounds are $B\,(\alpha/2)$-th and $B\,(1-\alpha/2)$-th ordered elements. For $B = 1000$ and $\alpha = 5\%$, these are the 25th and 975th ordered elements. The estimated $1-\alpha$ confidence interval is $\left[\widehat{\theta}^*_{(B\times(\alpha/2))}, \widehat{\theta}^*_{(B\times(1-\alpha/2))}\right]$.

What we did not discuss is whether the bootstrap is "correct". We need to show that for bootstrap standard errors,

$$\frac{SE}{\sigma/\sqrt{n}} \to_p 1 \tag{2}$$

and for the bootstrap percentile confidence interval,

$$\Pr\left(\theta \in \left[\widehat{\theta}^*_{(B\times(\alpha/2))}, \widehat{\theta}^*_{(B\times(1-\alpha/2))}\right]\right) \to 1-\alpha \tag{3}$$

as $n \to \infty$. This is a very difficult problem. Below we provide some discussion about why bootstrap "works". Bootstrap percentile confidence intervals often have more accurate coverage probabilities (i.e. closer to the nominal coverage probability $1-\alpha$) than the usual confidence intervals based on standard normal quantiles and estimated variance. The bootstrap percentile method is simple but it should not be abused. Loosely, it works in the sense that (3) is true, only if the estimator is asymptotically normal. Suppose we observe a random sample $X_1, ..., X_n$ from a uniform distribution on $[0, \theta]$, where $\theta > 0$ is unknown. $\widehat{\theta}_n = \max\{X_1, ..., X_n\}$ is a consistent estimator for θ and $n\left(\theta - \widehat{\theta}_n\right)$ converges in distribution to the exponential distribution. For this case, (3) fails. The bootstrap percentile method fails to give an asymptotically valid confidence interval.

How/Why Bootstrap Works?

Suppose we have an i.i.d. random sample $X_1, ..., X_n$ with CDF F_X. Suppose $S_n = \varphi_n(X_1, ..., X_n)$ is a statistic. Its distribution should depend on F_X:

$$F_{S_n}(x) = H(x \mid F_X) = \Pr(\varphi_n(X_1, ..., X_n) \leq x).$$

We know that the empirical CDF $\widehat{F}_X$ is a step function that jumps at each of $X_1, ..., X_n$ with size $1/n$. So $\widehat{F}_X$ is the CDF of a discrete random variable Z with $X_1, ..., X_n$ being its possible realizations and $1/n$ being the probability of any of $X_1, ..., X_n$ being selected:

$$\Pr(Z = X_k) = \frac{1}{n}, \text{ for each } k = 1, 2, ..., n.$$

A random observation from $X_1, ..., X_n$ is just a random variable that has the same distribution as Z. n observations randomly drawn with replacement from $X_1, ..., X_n$ are just a random sample from the distribution $\widehat{F}_X$. So each bootstrap sample is an i.i.d. random sample from $\widehat{F}_X$. Note that the "distribution" here should be interpreted as the conditional distribution given $X_1, ..., X_n$.

Let $X_1^*, ..., X_n^*$ be an i.i.d. random sample from $\widehat{F}_X$. Let $S_n^* = \varphi_n(X_1^*, ..., X_n^*)$. The conditional CDF given $X_1, ..., X_n$ of S_n^* is

$$H\left(x \mid \widehat{F}_X\right) = \Pr(\varphi_n(X_1^*, ..., X_n^*) \leq x \mid X_1, ..., X_n).$$

It seems reasonable to estimate $H(x \mid F_X)$ by $H\left(x \mid \widehat{F}_X\right)$ since $\widehat{F}_X$ is a very good estimate of F_X. Similar the variance $\text{Var}(S_n)$ should depend on F_X as well and it can be estimated by $\text{Var}(S_n^* \mid X_1, ..., X_n)$. The true (conditional) distribution of $X_1^*, ..., X_n^*$ is known. We can use computer simulations (known as Monte Carlo simulations) to compute $\text{Var}(S_n^* \mid X_1, ..., X_n)$. The computer draws B (very large) i.i.d. random samples from $\widehat{F}_X$ for us:

$$\begin{array}{cccc} X_1^{*(1)} & \cdots & X_n^{*(1)} & \text{i.i.d. } \sim \widehat{F}_X \\ X_1^{*(2)} & \cdots & X_n^{*(2)} & \text{i.i.d. } \sim \widehat{F}_X \\ \vdots & \ddots & \vdots & \vdots \\ X_1^{*(B)} & \cdots & X_n^{*(B)} & \text{i.i.d. } \sim \widehat{F}_X. \end{array}$$

These are just B independent bootstrap samples. Then,

$$\text{Var}\left(S_n^* \mid X_1, ..., X_n\right) \approx \frac{1}{B}\sum_{b=1}^{B}\left(\varphi_n\left(X_1^{*(b)}, ..., X_n^{*(b)}\right) - \overline{\varphi}_n\right)^2, \tag{4}$$

where $\overline{\varphi}_n = B^{-1}\sum_{b=1}^{B}\varphi_n\left(X_1^{*(b)}, ..., X_n^{*(b)}\right)$ is the (bootstrap) sample mean. Since B can be arbitrarily large, by WLLN, the right hand side of (4) should be very close to the left hand side.

What we put forward is just the intuition about how/why bootstrap works. The theoretical proof and also proof of the key results (2) and (3) are very difficult. Here is some further intuition.

Let $G(x) = \Pr\left(\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \leq x\right)$ be the distribution function of $\sqrt{n}\left(\widehat{\theta}_n - \theta\right)$. If we knew G, we could easily construct a confidence interval $\left[\widehat{\theta}_n - \frac{t_{1-\alpha/2}}{\sqrt{n}}, \widehat{\theta}_n - \frac{t_{\alpha/2}}{\sqrt{n}}\right]$,where t_α is the α-quantile of G: $t_\alpha = G^{-1}(\alpha)$. In reality, we do not know G and we can often show that G can be approximated by the distribution function of $N\left(0, \sigma^2\right)$. The normal approximation with $N\left(0, \sigma^2\right)$ requires that σ^2 can be estimated consistently. What bootstrap does is "alternative approximation". It suggests that the conditional distribution

$$\widehat{G}(x) = \Pr\left(\sqrt{n}\left(\widehat{\theta}_n^* - \widehat{\theta}\right) \leq x \mid X_1, ..., X_n\right),$$

where $\widehat{\theta}_n^*$ is the "bootstrap analogue" of $\widehat{\theta}_n$. $\widehat{\theta}_n^*$ is computed using the bootstrap random sample $X_1^*, ..., X_n^*$ but the same formula as $\widehat{\theta}_n$. The bootstrap random sample $X_1^*, ..., X_n^*$ are i.i.d. with CDF $\widehat{F}_X$. We can use the computer to generate as many samples as we want. $\widehat{G}$ is known to us since the distribution of the bootstrap sample is known. $\widehat{G}$ can be approximated by computer simulations. Indeed in many cases especially when $\sqrt{n}\left(\widehat{\theta}_n - \theta\right)$ is asymptotically normal, we have

$$\sup_{x\in\mathbb{R}}\left|\widehat{G}(x) - G(x)\right| \to_p 0.$$

So the estimation is consistent. But there are exceptions.

Bootstrap Refinement

If we have a plug-in estimator for σ and the estimator $\widehat{\sigma}_n$ is consistent, we have

$$T = \frac{\sqrt{n}\left(\widehat{\theta}_n - \theta\right)}{\widehat{\sigma}_n} \longrightarrow_d N(0,1).$$

Note that here $\widehat{\sigma}_n$ can be written as a function of the data and we know its function form. For each bootstrap sample $b = 1, ..., B$, we can calculate $\widehat{\sigma}_n^*$ using the bootstrap sample. For example, suppose $X_1, ..., X_n$ is an i.i.d. random sample with mean μ and variance σ^2. Let $\widehat{\mu}_n = n^{-1}\sum_{i=1}^{n} X_i$ and $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n}\left(X_i - \widehat{\mu}_n\right)^2$. We know

$$T = \frac{\sqrt{n}\left(\widehat{\mu}_n - \mu\right)}{\widehat{\sigma}_n} \longrightarrow_d N(0,1).$$

We can compute $\widehat{\sigma}_n^*$ as $\widehat{\sigma}_n^{*2} = n^{-1}\sum_{i=1}^{n}\left(X_i^* - \widehat{\mu}_n^*\right)^2$, with $\widehat{\mu}_n^* = n^{-1}\sum_{i=1}^{n} X_i^*$.

- Bootstrap-t

- Step 1: Draw B independent bootstrap samples. B can be as large as possible. We can take $B = 1000$.
- Step 2: Estimate θ and σ with each of the bootstrap samples, $\widehat{\theta}_n^{*(b)}$, $\widehat{\sigma}_n^{*(b)}$ for $b = 1, ..., B$ and the t-value for each bootstrap sample:

$$t_b^* = \frac{\sqrt{n}\left(\widehat{\theta}_n^{*(b)} - \widehat{\theta}_n\right)}{\widehat{\sigma}_n^{*(b)}}$$

- Step 3: Order the bootstrap replications of t such that $t_{(1)}^* \leq \cdots \leq t_{(B)}^*$.
- Step 4: The lower critical value $t_{\alpha/2}^*$ and the upper critical value $t_{1-\alpha/2}^*$ are then the $B \times (\alpha/2)$-th and $B \times (1 - \alpha/2)$-th ordered elements. For $B = 1000$ and $\alpha = 5\%$, these are the 25th and 975th ordered elements. The bootstrap lower and upper critical values generally differ in absolute values.

The bootstrap-t confidence interval is

$$\left[\widehat{\theta}_n + t_{2.5\%}^* \times \frac{\widehat{\sigma}_n}{\sqrt{n}}, \widehat{\theta}_n + t_{97.5\%}^* \times \frac{\widehat{\sigma}_n}{\sqrt{n}}\right].$$

A striking result is

$$\Pr\left(\theta \in \left[\widehat{\theta}_n + t_{2.5\%}^* \times \frac{\widehat{\sigma}_n}{\sqrt{n}}, \widehat{\theta}_n + t_{97.5\%}^* \times \frac{\widehat{\sigma}_n}{\sqrt{n}}\right]\right) = 95\% + O\left(n^{-3/2}\right)$$

compared with the confidence interval using the standard normal critical values

$$\Pr\left(\theta \in \left[\widehat{\theta}_n - 1.96 \times \frac{\widehat{\sigma}_n}{\sqrt{n}}, \widehat{\theta}_n + 1.96 \times \frac{\widehat{\sigma}_n}{\sqrt{n}}\right]\right) = 95\% + O\left(n^{-1}\right).$$

This is known as asymptotic "refinement" of bootstrap.

Residual Bootstrap and Wild Bootstrap

Consider the context of linear regression. Our observed data is $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ and we are interested in the regression coefficients:

$$Y_i = \alpha + \beta X_i + e_i.$$

In this case the nonparametric/empirical bootstrap we introduced works well, in the sense that the bootstrap standard errors are consistent and the bootstrap percentile confidence intervals have asymptotically correct coverage probabilities. Empirical bootstrap treats the pair (X, Y) as one object and each bootstrap sample consists of n independent observations drawn with replacement from the n observations $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$. There are popular alternatives to the empirical bootstrap. Bootstrap standard errors, percentile confidence intervals and bootstrap-t are carried out by following the same steps. The only thing that changes is how we resample to get the bootstrap samples.

Let $\widehat{e}_i = Y_i - \widehat{\alpha} - \widehat{\beta} X_i$, where $\left(\widehat{\alpha}, \widehat{\beta}\right)$ is the LS estimator. We draw n fitted residuals independently with replacement from $\widehat{e}_1, ..., \widehat{e}_n$. In other words, the bootstrap sample is an i.i.d. random sample

$\widehat{e}_1^*, ..., \widehat{e}_n^*$, where for each $i = 1, ..., n$,

$$\Pr\left(\widehat{e}_i^* = \widehat{e}_k\right) = \frac{1}{n}, \text{ for each } k = 1, 2, ..., n.$$

Now for each $i = 1, 2, ..., n$, let $X_i^* = X_i$ and $Y_i^* = \widehat{\alpha} + \widehat{\beta} X_i^* + \widehat{e}_i^*$. Note that the independent variables are the same in all bootstrap samples. This is known as the residual bootstrap.

For wild bootstrap, let $V_1, ..., V_n$ be n computer-generated independent random variables with mean zero that are also independent of the data. Now for each $i = 1, 2, ..., n$, let $\widehat{e}_i^* = V_i \times \widehat{e}_i$, $X_i^* = X_i$ and $Y_i^* = \widehat{\alpha} + \widehat{\beta} X_i^* + \widehat{e}_i^*$. The most popular distribution for V's is the following two-point "golden rule" distribution:

$$V_i = \begin{cases} -\left(\sqrt{5}-1\right)/2 & \text{with probability } \left(\sqrt{5}+1\right)/\left(2\sqrt{5}\right) \\ \left(\sqrt{5}+1\right)/2 & \text{with probability } \left(\sqrt{5}-1\right)/\left(2\sqrt{5}\right). \end{cases}$$

Its theoretical motivation was provided by Professor Enno Mammen in 1993.

Bootstrap Hypothesis Test

We now consider testing $\mathrm{H}_0 : \theta = \theta_0$. We can use any of the bootstrap-based confidence intervals and check if θ_0 is in the confidence interval. We simply reject H_0 if θ_0 fails to be an element of the bootstrap percentile confidence interval.

Since the t-statistic $T = \frac{\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)}{\widehat{\sigma}_n} \longrightarrow_d \mathrm{N}(0,1)$ under H_0. We use the standard normal distribution as approximation to the true distribution of T and define critical values based on standard normal quantile. Alternatively, we can do the following bootstrap-t test.

- Bootstrap-t test
  - Step 1: Draw B independent bootstrap samples. B can be as large as possible. We can take $B = 1000$.
  - Step 2: Estimate θ and σ with each of the bootstrap samples, $\widehat{\theta}_n^{*(b)}$, $\widehat{\sigma}_n^{*(b)}$ for $b = 1, ..., B$ and the t-value for each bootstrap sample:

$$t_b^* = \frac{\sqrt{n}\left(\widehat{\theta}_n^{*(b)} - \widehat{\theta}_n\right)}{\widehat{\sigma}_n^{*(b)}}$$

  - Step 3: Order the bootstrap replications of t such that $t_{(1)}^* \leq \cdots \leq t_{(B)}^*$.
  - Step 4: The lower critical value $t_{\alpha/2}^*$ and the upper critical value $t_{1-\alpha/2}^*$ are then the $B \times (\alpha/2)$-th and $B \times (1 - \alpha/2)$-th ordered elements. Reject H_0 if $T < t_{\alpha/2}^*$ or $T > t_{1-\alpha/2}^*$.

Caution: a common mistake is that in Step 2, one mistakenly computes

$$\frac{\sqrt{n}\left(\widehat{\theta}_n^{*(b)} - \theta_0\right)}{\widehat{\sigma}_n^{*(b)}}.$$

The test will have no power if we made this mistake. The distribution of the t-statistic $T = \frac{\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)}{\widehat{\sigma}_n}$

under H_1 is different from that under H_0. Under H_1, T is not centered:

$$T = \frac{\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)}{\widehat{\sigma}_n} = \frac{\sqrt{n}\left(\widehat{\theta}_n - \theta\right)}{\widehat{\sigma}_n} + \frac{\sqrt{n}\left(\theta - \theta_0\right)}{\widehat{\sigma}_n}.$$

An important guideline is that we should always approximate the distribution of T under H_0, i.e., the distribution of $\frac{\sqrt{n}\left(\widehat{\theta}_n - \theta\right)}{\widehat{\sigma}_n}$.