Lecture on Bootstrap

Yu-Chin Hsu (許育進)

Academia Sinica

November 2, 2015

- This lecture is based on Xiaoxia Shi's lecture note and my understanding of bootstrap.
- This is an introductory lecture to Bootstrap Method in that I won't provide any proofs.
- For further reading, please see
  Horowitz, J.L. (2001). "The Bootstrap", in J.J. Heckman and E. Leamer, eds, Handbook of Econometrics, vol. 5, Elsevier Science, B.V., p. 3159-3228,
  and references within.
- Xiaoxia Shi's lecture note is available at
  http://www.ssc.wisc.edu/∼ xshi/econ715/Lecture_10_bootstrap.pdf

Introduction

- Let $\mathcal{W} = (W_1, \ldots, W_n)$ denote an i.i.d. sample with distribution function F.
- Let θ_0 be the parameter of interest and $\hat{\theta}_n(W)$ denote the estimator based on $\mathcal{W}$.
  - For example, θ_0 can be the mean of W, $E[W]$, and $\hat{\theta}(\mathcal{W}) = n^{-1} \sum W_i$.
- To make inference or to construct confidence interval (CI) for θ_0, in general, we need to know the exact (or limiting) distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$.
- Most of the time, $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^2)$.
- Then given the availability of consistent estimator $\hat{\sigma}_n$ for σ, we can make inference and construct CI.

Introduction (cont'd)

- However, sometimes, (1) the exact form of σ is hard to obtain or (2) it is very complicated to construct consistent estimator $\hat{\sigma}$ for σ.
  - (1) can happen when the derivatives of the objective function of a maximum likelihood model are complicated.
  - (2) can happen when the $\hat{\sigma}$ involves nonparametric components, e.g., θ_0 is the medium of of F and $\hat{\theta}_n(\mathcal{W})$ is the sample medium. Then σ^2 will be $1/(4f^2(\theta_0))$ where $f(\cdot)$ is the pdf of F. Then to estimate $\hat{\sigma}$, one needs $\hat{f}(\hat{\theta}_n)$ which is a nonparametric estimator.
- Therefore, it is hard to make inference and to construct CI for θ_0.
- Bootstrap can be served as an alternate method for this purpose.

What is bootstrap?

- Shi: "Bootstrap is an alternative to asymptotic approximation for carrying out inference. The idea is to mimic the variation from drawing different samples from a population by the variation from redrawing samples from a sample."
- Horowitz: "The bootstrap is a method for estimating the distribution of an estimator or test statistic by resampling one's data or a model estimated from the data."
- Shi: "The name comes from the common English phrase "bootstrap" which alludes to "pulling oneself over the fence by pulling on ones own bootstrap", and means solving a problem without external help."

How does bootstrap work?

- Let $\mathcal{W}^* = (W_1^*, \ldots, W_n^*)$ denote an i.i.d. sample with distribution function F^*.
- Let θ_0^* be the parameter under F^* and $\hat{\theta}_n(W^*)$ denote the estimator based on $\mathcal{W}^*$.
- Basic Idea:
  - When F^* is close to F, then the distribution of $\sqrt{n}(\hat{\theta}^* - \theta_0^*)$ should be close to $\sqrt{n}(\hat{\theta} - \theta_0)$.
  - Therefore, if we can find an F^* (known to us) that is close to F (unknown), then we can approximate $\sqrt{n}(\hat{\theta} - \theta_0)$ (unknown) by $\sqrt{n}(\hat{\theta}^* - \theta_0^*)$ (known).
  - A natural choice of F^* is the empirical cdf $\widehat{F}_n$ since we can show that when sample size is large enough, $\widehat{F}_n$ is consistent for F. This leads to the "nonparametric bootstrap".

Confidence Interval

- What is the concept of a confidence interval (from a frequentist point of view)?
- Suppose $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, \sigma^2)$ and $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$.
- Then the two-sided 95% confidence interval of θ is:

$$\left(\hat{\theta} - \frac{1.96\hat{\sigma}}{\sqrt{n}}, \hat{\theta} + \frac{1.96\hat{\sigma}}{\sqrt{n}}\right).$$

- Why?

$$\begin{aligned}
& P(-1.96 < \sqrt{n}(\hat{\theta} - \theta_0)/\hat{\sigma} < 1.96) \approx 95\% \\
\Rightarrow & P(-1.96\hat{\sigma} < \sqrt{n}(\hat{\theta} - \theta_0) < 1.96\hat{\sigma}) \approx 95\% \\
\Rightarrow & P(-1.96\hat{\sigma}/\sqrt{n} < (\hat{\theta} - \theta_0) < 1.96\hat{\sigma}/\sqrt{n}) \approx 95\% \\
\Rightarrow & P(-1.96\hat{\sigma}/\sqrt{n} < (\theta_0 - \hat{\theta}) < 1.96\hat{\sigma}/\sqrt{n}) \approx 95\% \\
\Rightarrow & P(\hat{\theta} - 1.96\hat{\sigma}/\sqrt{n} < \theta_0 < \hat{\theta} + 1.96\hat{\sigma}/\sqrt{n}) \approx 95\%
\end{aligned}$$

Bootstrap Confidence Interval

- Suppose that the limiting distribution of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ is close to $\sqrt{n}(\hat{\theta} - \theta_0)$.
- Let's pretend that we know the 2.5% and 97.5% quantiles of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ for now and they are denoted as $q^*_{2.5}$ and $q^*_{97.5}$.
- Then the 95% CI is $(\hat{\theta} - q^*_{97.5}/\sqrt{n}, \hat{\theta} - q^*_{2.5}/\sqrt{n})$.
- Note that

$$
\begin{aligned}
& P(q^*_{2.5} < \sqrt{n}(\hat{\theta}^* - \hat{\theta}) < q^*_{97.5}) = 95\% \\
\Rightarrow & P(q^*_{2.5} < \sqrt{n}(\hat{\theta} - \theta_0) < q^*_{97.5}) \approx 95\% \\
\Rightarrow & P(q^*_{2.5}/\sqrt{n} < (\hat{\theta} - \theta_0) < q^*_{97.5}/\sqrt{n}) \approx 95\% \\
\Rightarrow & P(-q^*_{97.5}/\sqrt{n} < (\theta_0 - \hat{\theta}) < -q^*_{2.5}/\sqrt{n}) \approx 95\% \\
\Rightarrow & P(\hat{\theta} - q^*_{97.5}/\sqrt{n} < \theta_0 < \hat{\theta} - q^*_{2.5}/\sqrt{n}) \approx 95\%
\end{aligned}
$$

Remarks:

- In this example, we know that the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ is symmetric. Let α_{95} such that $P(\sqrt{n}|(\hat{\theta}^* - \hat{\theta})| < \alpha_{95}) = 95\%$.
- Then $P(\hat{\theta} - \alpha^*_{95}/\sqrt{n} < \theta_0 < \hat{\theta} - \alpha^*_{95}/\sqrt{n}) = 95\%$.
- Both CI's are asymptotically valid.
- In general, the second one can have higher-order improvement in that the converge rate of this CI converge to 95% at a faster rate than the first one. (Why?)
- In general, if the finite sample distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ is known to be skewed, then the first one might be a better one to use.

How to obtain those quantiles?

- So far, we pretend that $q^*_{2.5}$, $q^*_{97.5}$ and α^*_{95} are known.
- $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ is known as we pointed out, because we know W_i^* are drawn from $\widehat{F}$, the empirical CDF.
- Of course, the close form of the CDF of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ is still hard to get!
- Then this is where the well-known bootstrap simulations come into play.

How to obtain those quantiles? (Cont'd)

- We know that W_i^*'s are drawn from $\widehat{F}$ which is equivalent to randomly draw W_i^* from $\{W_1, \ldots, W_n\}$ with prob $1/n$.
- Therefore, a bootstrap sample $\{W_1^*, \ldots, W_n^*\}$ is formed from n random sample with replacement.
  - This step can be done by computer.
  - Generate $U[0, 1]$ random variables. Let u be a realization and we can have the index be k if $(k-1)/n < u \leq k/n$.

Bootstrap simulations:

1. We can use computer to draw $\{W^*_{1,b}, \ldots, W^*_{n,b}\}$ for $b = 1, \ldots, B$ and obtain $\hat{\theta}^*_b$.
2. Then the $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ can be further approximated by the empirical distribution of $\sqrt{n}(\hat{\theta}^*_b - \hat{\theta})$ from $b = 1, \ldots, B$.
3. Rank $\sqrt{n}(\hat{\theta}^*_{(b)} - \hat{\theta})$ in an ascending order such that $\sqrt{n}(\hat{\theta}^*_{(1)} - \hat{\theta}) \leq \sqrt{n}(\hat{\theta}^*_{(2)} - \hat{\theta}) \leq \ldots \leq \sqrt{n}(\hat{\theta}^*_{(B)} - \hat{\theta})$.
4. $q^*_{2.5}$ and $q^*_{97.5}$ can be approximated by $\hat{q}^*_{2.5} = \sqrt{n}(\hat{\theta}^*_{(\lfloor 2.5 * B \rfloor)} - \hat{\theta})$ and $\hat{q}^*_{97.5} = \sqrt{n}(\hat{\theta}^*_{(\lfloor 97.5 * B \rfloor)} - \hat{\theta})$, respectively, where $\lfloor c \rfloor$ denote the largest integer a such that $a \leq c$.
5. That is, if $B = 1000$, then $\hat{q}^*_{2.5} = \sqrt{n}(\hat{\theta}^*_{(25)} - \hat{\theta})$ and $\hat{q}^*_{97.5} = \sqrt{n}(\hat{\theta}^*_{(975)} - \hat{\theta})$, respectively.
6. $\hat{\alpha}^*_{95}$ is defined similarly except that the ranking is based on $\sqrt{n}|(\hat{\theta}^*_{(b)} - \hat{\theta})|$.

How to obtain those quantiles? (Cont'd)

- Note that this approximation can be as accurate as you please by setting B large enough.
- When B is too large, it might take too much time to compute. Therefore, there is a trade-off between accuracy and time.
- In general, setting $B = 700 \sim 1000$, the approximation can be good.
- Note that $\hat{q}^*_{97.5} = \sqrt{n}(\hat{\theta}^*_{(975)} - \hat{\theta})$. Therefore, the lower bound of the CI can be rewritten as

$$\hat{\theta} - \frac{\hat{q}^*_{97.5}}{\sqrt{n}} = \hat{\theta} - \frac{\sqrt{n}(\hat{\theta}^*_{(975)} - \hat{\theta})}{\sqrt{n}} = \hat{\theta} - (\hat{\theta}^*_{(975)} - \hat{\theta}).$$

- Similarly, the upper bound can be rewritten as $\hat{\theta} - (\hat{\theta}^*_{(25)} - \hat{\theta})$.

Hypothesis testing

- Let $W_i \sim N(\mu, 1)$. We want to test $H_0 : \ \mu = 1$ v.s. $H_0 : \ \mu \neq 1$ at 5% significance level.
- Test statistic: $\sqrt{n}(\hat{\mu}_n - 1)$ where $\hat{\mu}_n$ is the sample average.
- We would reject H_0 when $|\sqrt{n}(\hat{\mu}_n - 1)| > 1.96$
- Under $H_0 : \ \mu = 1$, we will falsely reject the null hypothesis 5% of the time.
- Under $H_1 : \ \mu \neq 1$, we will reject the null hypothesis with probability 1 asymptotically. (when $n \to \infty$)

A wrong bootstrap procedure!

- The following procedure is WRONG!
  1. Generate bootstrap samples: $\{W_{1,b}, \ldots, W_{n,b}\}$ for $b = 1, \ldots, B$, say $B = 1000$.
  2. Calculate $\sqrt{n}(\hat{\mu}_b^* - 1)$ and obtain $\tilde{q}_{2.5}^* = \sqrt{n}(\hat{\mu}_{(25)}^* - 1)$ and $\tilde{q}_{97.5}^* = \sqrt{n}(\hat{\mu}_{(975)}^* - 1)$.
  3. Reject H_0 when $\sqrt{n}(\hat{\mu}_n - 1) < \tilde{q}_{(25)}^*$ or $\sqrt{n}(\hat{\mu}_n - 1) > \tilde{q}_{(975)}^*$.
- To see why?

- Note that

$$\begin{aligned} & P(\sqrt{n}(\hat{\mu}_n - 1) < \tilde{q}^*_{(25)}) \\ =& P(\sqrt{n}(\hat{\mu}_n - 1) < \sqrt{n}(\hat{\mu}^*_{(25)} - 1)) \\ =& P(0 < \sqrt{n}(\hat{\mu}^*_{(25)} - \hat{\mu}_n)) \to 0. \end{aligned}$$

- Similarly,

$$\begin{aligned} & P(\sqrt{n}(\hat{\mu}_n - 1) > \hat{q}^*_{(975)}) \\ =& P(\sqrt{n}(\hat{\mu}_n - 1) > \sqrt{n}(\tilde{\mu}^*_{(975)} - 1)) \\ =& P(0 > \sqrt{n}(\hat{\mu}^*_{(975)} - \hat{\mu}_n)) \to 0. \end{aligned}$$

- Note that the previous two results hold no matter the true parameters are.
- Therefore, no matter under the null or under the alternative, the size or the power of such test is zero.

A Right way to do!

- The following procedure is correct!
  1. Generate bootstrap samples: $\{W_{1,b}, \ldots, W_{n,b}\}$ for $b = 1, \ldots, B$, say $B = 1000$.
  2. Calculate $\sqrt{n}(\hat{\mu}_b^* - \hat{\mu}_n)$ and obtain $\hat{q}_{2.5}^* = \sqrt{n}(\hat{\mu}_{(25)}^* - \hat{\mu}_n)$ and $\hat{q}_{97.5}^* = \sqrt{n}(\hat{\mu}_{(975)}^* - \hat{\mu}_n)$.
  3. Reject H_0 when $\sqrt{n}(\hat{\mu}_n - 1) < \hat{q}_{(25)}^*$ or $\sqrt{n}(\hat{\mu}_n - 1) > \hat{q}_{(975)}^*$.
- Why this is a valid procedure?

- Under the null hypothesis $H_0: \ \mu = 1$,

$$\begin{aligned} & P(\sqrt{n}(\hat{\mu}_n - 1) < \hat{q}^*_{(25)} \text{ or } \sqrt{n}(\hat{\mu}_n - 1) > \hat{q}^*_{(975)}) \\ =& 1 - P(\hat{q}^*_{(25)} < \sqrt{n}(\hat{\mu}_n - 1) < \hat{q}^*_{(975)}) \\ =& 1 - P(\hat{q}^*_{(25)} < \sqrt{n}(\hat{\mu}_n - \mu) < \hat{q}^*_{(975)}) \approx 0.05. \end{aligned}$$

- Under the alternative $H_1: \ \mu \neq 1$, we have $\sqrt{n}(\hat{\mu}_n - 1) \to \pm\infty$. Also, $\hat{q}^*_{(25)}$ and $\hat{q}^*_{(975)}$ are bounded in probability. Therefore,

$$P(\sqrt{n}(\hat{\mu}_n - 1) < \hat{q}^*_{(25)} \text{ or } \sqrt{n}(\hat{\mu}_n - 1) > \hat{q}^*_{(975)}) \to 1.$$

Remarks

- The key is to approximate the "null distribution" no matter we are under the null or under the alternative.
- In this case, the null distribution is $\sqrt{n}(\hat{\mu}_n - \mu)$ no matter the value of true parameter is.
- Therefore, we cannot just plug in the value that we want to test in the bootstrap repetitions.

Other uses of Bootstrap

Standard Error

- We can use bootstrap method to approximate the asymptotic standard error of an estimator, σ.
- As we mentioned, when constructing CI's or conducting hypothesis testing, we need a consistent estimator for σ.
- We can use bootstrap to obtain an consistent estimator:

$$\hat{\sigma}_n^* = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_b^* - \overline{\theta}^*\right)^2,$$

  where $\overline{\theta}^*$ is the sample average of $\hat{\theta}_b^*$'s.
- Then we can replace $\hat{\sigma}$ with $\hat{\sigma}_n^*$ in the previous cases.
- Shi's remark: To use bootstrap for standard error, the estimator under consideration must be asymptotically normal. Otherwise, the use of standard error itself is misguided.

Other uses of Bootstrap

Bias Correction

- We can use bootstrap to correct the bias of an estimator.
- The exact bias is $Bias(\hat{\theta}_n, \theta) = E[\hat{\theta}_n] - \theta$ and is unknown.
- The bootstrap estimator of the bias is:

$$\widehat{Bias}^*(\hat{\theta}_n, \theta) = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_b^* - \hat{\theta}_n.$$

- Then the bootstrap bias-corrected estimator for θ is

$$\hat{\theta}_{BC,n} = \hat{\theta}_n - \widehat{Bias}^*(\hat{\theta}_n, \theta) = 2\hat{\theta}_n - \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_b^*.$$

- Shi's remark: Bias correction usually increases the variance because the bias is estimated. (This causes a trade-off between bias and variance.) Therefore it should not be used indiscriminately.

Higher-order improvements of the Bootstrap

- This part is rather theoretical, so we will skip it. Please see Shi's note for more discussions.

Bootstrap for Regression Models

- The regression model we consider is

$$Y_i = X_i\beta + U_i, \quad \text{for } i = 1, \ldots, n,$$

  where $W_i = (Y_i, X_i')$ is iid with F.
- Let $\hat{\beta}_n$ denote the OLS estimator for β such that

$$\hat{\beta}_n = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i.$$

- Under regularity conditions, $\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{D} N(0, V)$ where

$$V = E[XX']^{-1} E[U^2 XX'] E[XX']^{-1}.$$

Bootstrap for Regression Models (Cont'd)

- The nonparametric bootstrap works here.
- Bootstrap sample are form by the pairs of $W_i = (Y_i, X_i')'$.
- That is, a bootstrap sample $\{(Y_i^*, X_i^*)\}_{i=1}^n$ is a random sample with replacement from $\{(Y_i, X_i)\}_{i=1}^n$.
- $\hat{\beta}_n^*$ is calculated in the same way as $\hat{\beta}_n$:

$$\hat{\beta}_n^* = \left(\frac{1}{n}\sum_{i=1}^n X_i^* X_i^{*\prime}\right)^{-1} \frac{1}{n}\sum_{i=1}^n X_i^* Y_i^*.$$

- Then, results similar to what we discussed before would hold in this case under suitable conditions.

Wild Bootstrap for Regression Models

- In OLS, we have $Y_i = X_i\hat{\beta}_n + \hat{e}_i$ where $\hat{e}_i$'s are the residuals.
- Let U_i^b's denote iid pseudo random variables with mean 0 and variance 1.
- Let the b-th bootstrap sample be

$$Y_{b,i}^* = X_i\hat{\beta}_n + \hat{e}_i \cdot U_i^b,$$

  and regressors are X_i's.
- Then the $\hat{\beta}_b^*$ is

$$\hat{\beta}_b^* = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^n X_i Y_{b,i}^* = \hat{\beta}_n + \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^n X_i\hat{e}_i \cdot U_i^b.$$

- Then we can show that $\sqrt{n}(\hat{\beta}_n^* - \hat{\beta}_n)$ can approximate $\sqrt{n}(\hat{\beta}_n - \hat{\beta})$ well.
- This is a residual-based bootstrap and this only works for OLS.

Bootstrap method for weakly dependent data

- The bootstrap method we discuss above only works in iid framework.
- For weakly dependent data, the dependence among observations plays an important role in the asymptotics.
- Doing the nonparametric bootstrap above will not work because it will break down the dependence.
- We need a method that can mimic the dependence structure.

Blockwise Bootstrap

- Instead of resample an observation, we resample a bunch of observations together.
- To be specific, let the block size be k and the sample size be T. Then we have $T - k + 1$ blocks:

$$
\begin{aligned}
&(W_1, W_2, \ldots, W_k) \\
&(W_2, W_3, \ldots, W_{k+1}) \\
&\quad \vdots \\
&(W_{T-k+1}, \ldots, W_T).
\end{aligned}
$$

- To form a bootstrap sample,
  1. we randomly select m blocks (with replacement) such that $m \cdot k \geq T$ and $(m-1) \cdot k < T$.
  2. laying them end-to-end in the order sampled.
  3. Drop the last $m \cdot k - T$ observations from the last sampled block so that the sample size of the bootstrap sample is equal to T.

Blockwise Bootstrap (Cont'd)

- For this method to work asymptotically, we require the block size $k \to \infty$, but $k/T \to 0$ at a suitable rate.
- Why this method would work?
- Why $k \to \infty$, but $k/T \to 0$??

Conclusion!