The bootstrap

Giselle Montamat

Harvard University

Spring 2020

The bootstrap

Main idea: an approach to inference that instead of relying on normal asymptotic approximation to the true distribution of a statistic, finds an estimate of this distribution that is based on resampling from the sampled data.

$$\text{Data: } D_1, ..., D_n \overset{i.i.d}{\sim} F_0$$

$$\text{Statistic: } S_n = s_n(D_1, ..., D_n)$$

$$\text{True distribution: } S_n \sim P_{F_0}(S_n \leq s)$$

For example: $S_n = \hat{\theta}$; $S_n = \sqrt{n}(\hat{\theta} - \theta_0)$; $S_n = \frac{\hat{\theta}}{se(\hat{\theta})}$

Note 1: $P_{F_0}(S_n \leq s)$ is a function $G_n(s, F_0)$. Keep this in mind but I'll be using the $P_{F_0}(S_n \leq s)$ notation to remind ourselves that it is the distribution function of the statistic.

Note 2: in these notes, we assume i.i.d data but see comments on what changes if data clustered.

The bootstrap

- Asymptotic normality approach

$$P_{F_0}(S_n \leq s) \approx \lim_{n \to \infty} P_{F_0}(S_n \leq s) = \underbrace{P_{F_0}(S_\infty \leq s)}_{\substack{\text{Found analytically,} \\ \text{normal}}}$$

- Bootstrap approach

$$P_{F_0}(S_n \leq s) \approx \underbrace{P_{\hat{F}}(S_n \leq s)}_{\substack{\text{Distribution of } S_n \\ \text{when } D \sim \hat{F}. \\ \text{Estimated with} \\ \text{simulation approach} \\ \text{by drawing from } \hat{F}}}$$

The bootstrap

General algorithm:

1. For each $b = 1, ..., B$, where B is large (say 10,000):
   1. Generate a bootstrap sample of size n, $D^b = (D_1^b, ..., D_n^b)$, by drawing from $\hat{F}$, an estimate of F_0.
   2. Compute S_n^b for this bootstrap sample.
2. Use the computed $(S_n^1, ..., S_n^B)$ to get an empirical distribution of S_n^b and use this as an approx of $P_{F_0}(S_n \leq s)$:

$$P_{F_0}(S_n \leq s) \approx P_{\hat{F}}(S_n \leq s) \underset{(*)}{\approx} \frac{1}{B} \sum_{b=1}^{B} 1(S_n^b \leq t)$$

(*) Can show:

$$\frac{1}{B} \sum_{b=1}^{B} 1(S_n^b \leq t) \xrightarrow{B \to \infty} P_{\hat{F}}(S_n \leq s)$$

The bootstrap

Different bootstrap approaches suggest different $\hat{F}$:

1. "Infeasible" bootstrap (just a theoretical exercise)
2. Non-parametric bootstrap (most common - Isaiah: use this unless good reason not to)
3. Parametric bootstrap
4. Residual bootstrap
5. Wild bootstrap

"Infeasible" bootstrap

Assumes that we know F_0 (note: this is a theoretical exercise...if F_0 is known, then estimation and inference not needed...).

1. For each $b = 1, ..., B$, where B is large (say 10,000):
   1. Generate a bootstrap sample of size n, $D^b = (D_1^b, ..., D_n^b)$, by drawing from F_0.
   2. Compute S_n^b for this bootstrap sample.
2. Use the computed $(S_n^1, ..., S_n^B)$ to get an empirical distribution of S_n^b and use this as an approx of $P_{F_0}(S_n \leq s)$:

$$P_{F_0}(S_n \leq s) \approx \frac{1}{B} \sum_{b=1}^{B} 1(S_n^b \leq t)$$

Glivenko-Cantelli Theorem (empirical distributions converge to true distributions when iid sample grows):

$$\frac{1}{B} \sum_{b=1}^{B} 1(S_n^b \leq t) \xrightarrow{B \to \infty} P_{F_0}(S_n \leq s)$$

Non-parametric bootstrap

Uses the empirical distribution of D as $\hat{F}$:

$$\hat{F}(d) = \frac{1}{n}\sum_{i=1}^{n} 1(D_i \leq d)$$

1. For each $b = 1, ..., B$, where B is large (say 10,000):
    1. Generate a bootstrap sample of size $n(*)$, $D^b = (D_1^b, ..., D_n^b)$, by drawing from $\hat{F}$. **In practice, this can be done by randomly sampling with replacement from $D = (D_1, ..., D_n)$.**
    2. Compute S_n^b for this bootstrap sample.
2. Use the computed $(S_n^1, ..., S_n^B)$ to get an empirical distribution of S_n^b and use this as an approx of $P_{F_0}(S_n \leq s)$:

$$P_{F_0}(S_n \leq s) \approx P_{\hat{F}}(S_n \leq s) \approx \frac{1}{B}\sum_{b=1}^{B} 1(S_n^b \leq t)$$

(*)Note: if clustered data with clusters of different size, sample size of each bootstrap sample can vary.

Parametric bootstrap

Assumes that F_0 is within a family of distributions: $F_0 \in \{F(., \theta) : \theta \in \Theta\}$, so there's a θ_0 such that $F_0(d) = F(d, \theta_0)$. Thus, approximate F_0 with:

$$\hat{F} = F(., \hat{\theta})$$

$$\hat{\theta} \xrightarrow{p} \theta_0$$

1. For each $b = 1, ..., B$, where B is large (say 10,000):
    1. Generate a bootstrap sample of size n, $D^b = (D_1^b, ..., D_n^b)$, by drawing from $\hat{F}$(*)
    2. Compute S_n^b for this bootstrap sample.
2. Use the computed $(S_n^1, ..., S_n^B)$ to get an empirical distribution of S_n^b and use this as an approx of $P_{F_0}(S_n \leq s)$:

$$P_{F_0}(S_n \leq s) \approx P_{\hat{F}}(S_n \leq s) \approx \frac{1}{B} \sum_{b=1}^{B} 1(S_n^b \leq t)$$

(*)Note: if clustered data, draw at the cluster level.

Residual bootstrap

Assumes that $Y_i = h(X_i, \theta) + \epsilon_i$, $E[\epsilon_i | X_i] = 0$.

1. Compute estimate $\hat{\theta}$ by OLS/NLS and residuals $\hat{\epsilon}_i = Y_i - h(X_i, \hat{\theta})$.
2. For each $b = 1, ..., B$, where B is large (say 10,000):
    1. Generate a bootstrap sample of size n, $\hat{\epsilon}^b = (\hat{\epsilon}_1^b, ..., \hat{\epsilon}_n^b)$, by randomly sampling with replacement from $(\hat{\epsilon}_1, ..., \hat{\epsilon}_n)$.
    2. Generate a bootstrap sample of size n, $X^b = (X_1^b, ..., X_n^b)$, by randomly sampling with replacement from $(X_1, ..., X_n)$.
    3. Generate a bootstrap sample of size n, $Y^b = (Y_1^b, ..., Y_n^b)$ by computing $Y_i^b = h(X_i^b, \hat{\theta}) + \epsilon_i^b$ (note that because the previous two steps are independent from each other, we're implicitly imposing homoskedasticity. Also, if clustered data, clusters must be same size).
    4. Compute S_n^b for this bootstrap sample (Y^b, X^b). For example: $S_n^b = \hat{\theta}^b$, or $S_n^b = \sqrt{n}(\hat{\theta}^b - \hat{\theta})$.
3. Use the computed $(S_n^1, ..., S_n^B)$ to get an empirical distribution of S_n^b and use this as an approx of $P_{F_0}(S_n \leq s)$:

$$P_{F_0}(S_n \leq s) \approx P_{\hat{F}}(S_n \leq s) \approx \frac{1}{B} \sum_{b=1}^{B} 1(S_n^b \leq t)$$

Wild bootstrap

Assumes that $Y_i = h(X_i, \theta) + \epsilon_i$, $E[\epsilon_i | X_i] = 0$.

1. Compute estimate $\hat{\theta}$ by OLS/NLS and residuals $\hat{\epsilon}_i = Y_i - h(X_i, \hat{\theta})$.
2. For each $b = 1, ..., B$, where B is large (say 10,000):
   1. (Generate a bootstrap sample of size n, $(X^b, \hat{\epsilon}^b) = ((X_1^b, \hat{\epsilon}_1^b), ..., (X_n^b, \hat{\epsilon}_n^b))$, by randomly sampling with replacement from $((X_1, \hat{\epsilon}_1), ..., (X_n, \hat{\epsilon}_n))$.
   2. Generate a bootstrap sample of size n, $Y^b = (Y_1^b, ..., Y_n^b)$ by computing $Y_i^b = h(X_i^b, \hat{\theta}) + V_i^b * \epsilon_i^b$ where V_i^b is drawn from a distribution that takes values -1 and 1 with equal probability (it flips the sign of the residual). (If clustered data, V_i^b is the same for all i of same cluster).
   3. Compute S_n^b for this bootstrap sample (Y^b, X^b). For example: $S_n^b = \hat{\theta}^b$, or $S_n^b = \sqrt{n}(\hat{\theta}^b - \hat{\theta})$.
3. Use the computed $(S_n^1, ..., S_n^B)$ to get an empirical distribution of S_n^b and use this as an approx of $P_{F_0}(S_n \leq s)$:

$$P_{F_0}(S_n \leq s) \approx P_{\hat{F}}(S_n \leq s) \approx \frac{1}{B} \sum_{b=1}^{B} 1(S_n^b \leq t)$$

The bootstrap

Suppose $S_n = \hat{\theta}$. From simulation procedure, we got $(\hat{\theta}^1, ..., \hat{\theta}^B)$. Now we want to do inference; specifically, suppose we want 95% confidence intervals. Different options:

1. Rely on asymptotic normal approximation of the distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ but use bootstrap standard error of $\hat{\theta}$ instead of formula for asymptotic variance:

$$CI = \left[\hat{\theta} - 1.96 \times se_{boot}(\hat{\theta}) \; ; \; \hat{\theta} + 1.96 \times se_{boot}(\hat{\theta})\right]$$

$$se_{boot}(\hat{\theta}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^b - \bar{\hat{\theta}}\right)^2}$$

$$\bar{\hat{\theta}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^b$$

The bootstrap

2. Find the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the empirical distribution of $(\hat{\theta}^1, ..., \hat{\theta}^B)$, for $\alpha = 0.05$.

$$CI = \left[q_{\frac{\alpha}{2}} \ ; \ q_{1-\frac{\alpha}{2}}\right]$$

Called "percentile bootstrap interval".

The bootstrap

3. Take $S_n = \frac{\hat{\theta}-\theta_0}{\hat{\sigma}}$. From simulation procedure, obtain $(S_n^1, ..., S_n^B)$ where $S_n^b = \frac{\hat{\theta}^b-\hat{\theta}}{\hat{\sigma}^b}$ (note that in this case, we're obtaining $\hat{\sigma}^b$ for each bootstrap sample using the asymptotic formula for the variance).

   Find the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the empirical distribution of $(S_n^1, ..., S_n^B)$, for $\alpha = 0.05$ and use these as critical values for constructing the interval.

$$CI = \left[\hat{\theta} - q_{1-\frac{\alpha}{2}} \times \hat{\sigma} \; ; \; \hat{\theta} - q_{\frac{\alpha}{2}} \times \hat{\sigma}\right]$$

   Called "bootstrap t interval".

The bootstrap

Exercise: In problem set 6, exercise 2 you were asked to find the percentile interval and t-interval based on the non-parametric bootstrap for a GMM-IV estimator.

$$Y_i = T_i\beta + X_i'\delta + e_i$$

Moment conditions:

$$E[g(D_i;\gamma)] = E\left[\begin{pmatrix} Z_i \\ X_i \end{pmatrix} e_i\right] = E\left[\begin{pmatrix} Z_i \\ X_i \end{pmatrix} Y_i - \begin{pmatrix} Z_i \\ X_i \end{pmatrix}\begin{pmatrix} T_i \\ X_i \end{pmatrix}'\begin{pmatrix} \beta \\ \delta \end{pmatrix}\right] = 0$$

System is just-identified so can find $\hat{\beta}$ and $\hat{\delta}$ directly from sample analog of these moments:

$$\begin{pmatrix} \hat{\beta} \\ \hat{\delta} \end{pmatrix} = \left[\frac{1}{n}\sum_i \begin{pmatrix} Z_i \\ X_i \end{pmatrix}\begin{pmatrix} T_i \\ X_i \end{pmatrix}'\right]^{-1}\left[\frac{1}{n}\sum_i \begin{pmatrix} Z_i \\ X_i \end{pmatrix} Y_i\right]$$

The bootstrap

Formula for asymptotic variance of GMM estimator:

$$Var(\hat{\gamma}) = (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}$$

If system is just-identified, formula simplifies (G and W invertible):

$$Var(\hat{\gamma}) = G^{-1}\Omega(G^{-1})'$$

$$G = E\left[\frac{\partial}{\partial\gamma}g(D_i,\gamma)\right] = -E\left[\begin{pmatrix} Z_i \\ X_i \end{pmatrix}\begin{pmatrix} T_i \\ X_i \end{pmatrix}'\right] \Rightarrow \hat{G} = -\frac{1}{n}\sum_i \begin{pmatrix} Z_i \\ X_i \end{pmatrix}\begin{pmatrix} T_i \\ X_i \end{pmatrix}'$$

$$\Omega = Var(g(D_i,\gamma)) = E[g(D_i,\gamma)g(D_i,\gamma)']$$

$$\hat{\Omega}_{iid} = \frac{1}{n}\sum_i \left[\begin{pmatrix} Z_i \\ X_i \end{pmatrix}\hat{e}_i\hat{e}_i'\begin{pmatrix} Z_i \\ X_i \end{pmatrix}'\right]$$

$$\hat{\Omega}_{clus} = \frac{1}{n}\sum_c \left[\sum_{i\in I(c)}\begin{pmatrix} Z_i \\ X_i \end{pmatrix}\hat{e}_i\right]\left[\sum_{i\in I(c)}\begin{pmatrix} Z_i \\ X_i \end{pmatrix}\hat{e}\right]'$$

The bootstrap

- Asymptotic standard error of $\hat{\beta}$ and 95% CI:

$$se(\hat{\beta}) = \sqrt{\frac{[\hat{G}^{-1}\hat{\Omega}(\hat{G}^{-1})']_{11}}{n}}$$

$$CI = \left[\hat{\beta} - 1.96 \times se(\hat{\beta}) \;;\; \hat{\beta} + 1.96 \times se(\hat{\beta})\right]$$

- 95% percentile bootstrap interval (based on non-parametric bootstrap):
  Find the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the empirical distribution of $(\hat{\beta}^1, ..., \hat{\beta}^B)$, for $\alpha = 0.05$.

$$CI = \left[q_{\frac{\alpha}{2}} \;;\; q_{1-\frac{\alpha}{2}}\right]$$

Note: difference between clustered and iid data is how you resample from your data in your boostrap algorithm.

The bootstrap

- 95% t bootstrap interval (based on non-parametric bootstrap): Find the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the empirical distribution of $\left(\frac{\hat{\beta}^1-\hat{\beta}}{se^1(\hat{\beta})}, ..., \frac{\hat{\beta}^B-\hat{\beta}}{se^B(\hat{\beta})}\right)$, for $\alpha = 0.05$ and use these as critical values for constructing the interval.

$$CI = \left[\hat{\beta} - q_{1-\frac{\alpha}{2}} \times se(\hat{\beta})\ ;\ \hat{\beta} - q_{\frac{\alpha}{2}} \times se(\hat{\beta})\right]$$

Note: difference between clustered and iid data is how you resample from your data in your boostrap algorithm and what formula for s.e. you use.

The bootstrap

Bootstrap v asymptotic approaches to inference

- Both rely on asymptotics. Bootstrap distribution is a "good" approximation of true distribution but this is a convergence statement:

$$P_{F_0}(S_n \leq s) \approx P_{\hat{F}}(S_n \leq s) \text{ since: } \underset{s}{sup}|P_{\hat{F}}(S_n \leq s) - P_{F_0}(S_n \leq s)| \xrightarrow{p} 0$$

  Moreover, this convergence result usually relies on another convergence result:

$$P_{F_0}(S_n \leq s) \xrightarrow{p} P_{F_0}(S_\infty \leq s)$$

  So both rely on the same asymptotic result (ie, bootstrap "works" when asymptotic normality holds).

- For smaller samples, bootstrap better. Theory can show that bootstrap CI coverage converges to true finite sample CI coverage for some statistics (eg, t stats); this also seems be true more generally but we don't know how to prove it.