STAT 740: Bootstrap

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 740: Statistical Computing

Bootstrap etymology

To "*lift youself up by your bootstraps*" means to pull yourself out of a difficult situation through sheer effort, without outside help.

Here, this refers to using the *data itself* to help us estimate a statistic's sampling distribution.

The bootstrap is computationally demanding but easy to code, and very useful when

- Asymptotics don't work; e.g. non-standard or non-regular models, small sample sizes, richly parameterized, etc.
- Asymptotics are too difficult to carry out.

Key idea: Sampling variability in the statistic is approximated very well by the sample itself! "Proof" of bootstrap approximations relies on von Mises differentiable functionals or Edgeworth expansions...we wont worry about that here. Basic idea (like kernel estimation) has been around a long time without formal "proofs."

Statistic $\hat{\theta}(\mathbf{x}_{1:n})$

Let data $\mathbf{x}_{1:n} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, where $\mathbf{x}_i \in \mathbb{R}^k$, be

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} F.$$

Let $\hat{\theta}(\mathbf{x}_{1:n})$ be a statistic that estimates a population parameter θ, e.g. the mean, IQR, a quantile, etc. θ can also be (parametric) model parameters or a function of parameters.

The empirical measure based on data $\mathbf{x}_{1:n}$ is $F_n = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$, where $\delta_{\mathbf{x}_i}$ is point mass or "Dirac measure" at $\mathbf{x}_i$. Corresponding empirical c.d.f. is

$$F_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} I\{x_{i1} \leq x_1, \ldots, x_{ik} \leq x_k\}.$$

Note that $F_n(\mathbf{x}) \overset{a.s.}{\to} F(\mathbf{x})$.

Sampling distribution of $\hat{\theta}(\mathbf{x}_{1:n})$

The statistic $\hat{\theta}(\mathbf{x}_{1:n})$ is a random vector and has a distribution, called the *sampling distribution* induced by F.

The sampling distribution is important; it's how we perform inference on θ, e.g. hypothesis tests, confidence intervals, etc.

A Monte Carlo estimate of the (unknown!) sampling distribution is obtained by repeatedly taking independent samples of size n and forming the statistic

$$\mathbf{x}_{m,1}, \ldots, \mathbf{x}_{m,n} \overset{iid}{\sim} F, \quad \hat{\theta}_m = \hat{\theta}(\mathbf{x}_{m,1:n}), \quad m = 1, \ldots, M.$$

The Monte Carlo sample $\hat{\theta}_1, \ldots, \hat{\theta}_M$ can be used to make a histogram of, estimate quantiles from, or form integrals against the sampling distribution of $\hat{\theta}(\mathbf{x}_{1:n})$.

Approximate sampling distribution of $\hat{\theta}(\mathbf{x}_{1:n})$

Unfortunately, we don't know F! However, noting that that $F_n(\mathbf{x}) \overset{a.s.}{\to} F(\mathbf{x})$, an approximate Monte Carlo estimate of the sampling distribution is obtained by repeatedly taking independent samples of size n from F_n and forming the statistic

$$\mathbf{x}_{m,1}, \ldots, \mathbf{x}_{m,n} \overset{iid}{\sim} F_n, \quad \hat{\boldsymbol{\theta}}_m = \hat{\boldsymbol{\theta}}(\mathbf{x}_{m,1:n}), \quad m = 1, \ldots, M.$$

That's it! We simply replace F by F_n...it's like *magic*.

The key property is that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are *iid* from F. So if we have, say, regression data $\{(\mathbf{z}_i, y_i)\}_{i=1}^n$, then $\mathbf{x}_i = (\mathbf{z}_i, y_i)$. If we have survival data w/ risk factors $\mathbf{x}_i = (\mathbf{z}_i, y_i, \delta_i)$, etc.

What can we do with our Monte Carlo sample?

- Estimate the standard error $\mathrm{SE}(\hat{\theta}_j)$.
- Form confidence intervals for θ_j from $\hat{\boldsymbol{\theta}}_{1,j}, \ldots, \hat{\boldsymbol{\theta}}_{M,j}$ by simply taking quantiles. In the `boot` packages this is `method="perc"`.
- Test, e.g., $H_0 : \theta_j = b$ based on bootstrap CI.
- Examine the sampling distribution of $\hat{\theta}_j(\mathbf{x}_{1:n})$ via histograms or kernel-smoothed densities.
- Reduce estimation bias by considering the bootstrap sampling approximation to $\hat{\theta}_j - \theta_j$. Tim thinks this is silly, as a general approach (board).

Examples: quantiles; logistic regression standard errors & CIs.

Quantile CIs

Obtaining a CI for a population quantile is challenging. The estimator itself is easy; let

$$x_1, \ldots, x_n \overset{iid}{\sim} F,$$

and let F_n be the e.c.d.f. Define $F_n^{-1}(p) = \inf\{x : F_n(x) \geq p\}$. Then the $p(100)$th quantile is $q_n = F_n^{-1}(p) = x_{(\lceil np \rceil)}$. Note: *this is not what R uses!* Then

$$q_n \overset{\bullet}{\sim} N\left(q, \frac{p(1-p)}{nf(q)^2}\right)$$

can be used to obtain a large-sample CI for q from estimating f by a kernel-smoothed version $\hat{f}$.

Alternatively, we can use the bootstrap! **Example.**

Logistic regression w/ small sample size

Exact logistic regression is useful in small samples where large-sample asymptotics fail. An alternative approach to obtaining standard errors is simply bootstrapping the coefficients. Let the data be $\mathbf{x}_{1:n} = \{(\mathbf{z}_i, y_i)\}_{i=1}^n$. Simply sample n pairs $(\mathbf{z}_i, y_i)$ from $\mathbf{x}_{1:n}$ with replacement, computing the MLE for each sample:

$$\hat{\boldsymbol{\beta}}^1, \ldots, \hat{\boldsymbol{\beta}}^M.$$

The estimated SE for each $\hat{\beta}_j$ is simply the standard deviation

$$\widehat{SE}(\hat{\beta}_j) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\beta}_j^m - \frac{1}{M}\sum_{s=1}^{N}\hat{\beta}_j^s\right)^2}.$$

Note: Bayesian approach also natural.

Logistic regression w/ small sample size

In small sample sizes, or with perfectly predictive data, MLEs may not be unique or exist. This happens when (a) all of the responses are zeros or all ones, or (b) the zeros and ones can be perfectly separated by a hyperplane (quasi or complete separation).

One method for obtaining estimates (no longer MLEs) is to simply *use a prior on* $\boldsymbol{\beta}$, the most natural being Jeffreys' prior:

$$\pi(\boldsymbol{\beta}) \propto |\mathbf{X}'\mathbf{M}\mathbf{X}|^{1/2}, \quad \mathbf{M} = \text{diag}\left(\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{(1+e^{\mathbf{x}_i'\boldsymbol{\beta}})^2}\right).$$

Frequentists call the posterior mode under this prior a *penalized likelihood MLE*. They would rather use the term "penalized likelihood" than "prior" or "Bayes." Originally due to Firth (1993).

Asymptotic CIs using Firth's method can be terrible; bootstrap a natural approach here. Firth's method also necessary because a bootstrapped sample might have all zeros or all ones!
Example: Challenger data.

Parametric bootstrap

The bootstrap just discussed applies to any situation – parametric or not – and is referred to as the "nonparametric bootstrap."

The parametric bootstrap is most useful for estimating p-values for testing hypotheses in *parametric models* in the presence of "nuisance parameters."

Say $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} F_{\boldsymbol{\theta}},$$

and we want to test $H_0 : \boldsymbol{\theta}_2 = \mathbf{b}$, or even the nonlinear $H_0 : \mathbf{g}(\boldsymbol{\theta}_2) = \mathbf{b}$.

Let $\hat{\boldsymbol{\theta}}_{10}$ be the MLE of $\boldsymbol{\theta}_1$ under the constraint H_0 on $\boldsymbol{\theta}_2$ and let $W = \mathbf{h}(\mathbf{x}_{1:n})$ be a test-statistic (not a function of $\boldsymbol{\theta}$) to test H_0 (bigger $W \Rightarrow H_0$ unlikely).

Parametric bootstrap

Key observation: MLEs are functions of sufficient statistics, so sampling from $F_{\hat{\boldsymbol{\theta}}_{10}}$ is equivalent to conditioning on the sufficient statistics under H_0. Related to score test; only fit reduced model!

A bootstrapped p-value takes

$$\mathbf{x}_{m,1}, \ldots, \mathbf{x}_{m,n} \overset{iid}{\sim} F_{\hat{\boldsymbol{\theta}}_{10}}, \quad W_m = \mathbf{h}(\mathbf{x}_{m,1:n}), \quad m = 1, \ldots, M,$$

and $p = \frac{1}{M} \sum_{m=1}^{M} I\{W_m \geq W\}$.

Since the sampling distribution relies on the underlying parametric model, a parametric bootstrap may be sensitive to parametric assumptions, unlike the nonparametric bootstrap.

Examples: independence in $r \times c$ tables, multivariate example.

Testing independence in $r \times c$ table

Let $\mathbf{n} = \{n_{ij}\}$ be the cell counts for $i = 1, \ldots, r$ and $j = 1, \ldots, c$. Let $\pi_{ij} = P(X = i, Y = j)$. Under $H_0 : X \perp Y$, $\hat{\pi}_{ij} = \frac{n_{i+}}{n_{++}} \frac{n_{+j}}{n_{++}}$. The Pearson test statistic is

$$W(\mathbf{n}) = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}},$$

where $\hat{\mu}_{ij} = \frac{n_{i+} n_{+j}}{n_{++}}$.

Parametric bootstrap simply repeatedly samples a multinomial over a table $\mathbf{n}_m \sim \text{mult}(n, \hat{\boldsymbol{\pi}})$, where $\boldsymbol{\pi} = \{\hat{\pi}_{ij}\}$ is computed under H_0 as described above. Then

$$p = \frac{1}{M} \sum_{m=1}^{M} I\{W(\mathbf{n}_m) \geq W(\mathbf{n})\}.$$

Testing independence in data $\mathbf{x}_i$

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. A test of interest is that the elements of each $\mathbf{x}_i$ are independent, $H_0 : \boldsymbol{\Sigma} = \kappa \mathbf{I}_p$. Mardia, Kent, and Bibby (1979, Chapter 5) derive the LRT

$$L = np \log \left\{ \frac{\frac{1}{p}\text{tr}\mathbf{S}}{|\mathbf{S}|^{1/p}} \right\}.$$

Under H_0, the MLE's are $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and $\hat{\kappa} = \frac{1}{p}\text{tr}\mathbf{S}$. A bootstrapped p-value simulates M bootstrap samples

$$\mathbf{x}_1^m, \ldots, \mathbf{x}_n^m \overset{iid}{\sim} N_p(\hat{\boldsymbol{\mu}}, \hat{\kappa}\mathbf{I}_p),$$

and forms test statistics $L_m = np \log \left\{ \frac{\frac{1}{p}\text{tr}\mathbf{S}_m}{|\mathbf{S}_m|^{1/p}} \right\}$. The p-value is

$$p = \frac{1}{M} \sum_{m=1}^{M} I\{L_m \geq L\}.$$

Comments

- Bootstrap is not a panacea; breaks down for small samples and/or dependent data. There are various fixes, e.g. smoothed bootstrap, blocked bootstrap, etc.
- Very useful when asympotics are either inaccurate or too difficult to obtain.
- Can be more robust than large sample results.
- Computationally intensive, especially, e.g. repeated bootstraps, etc.
- Bayesian approach also natural; posterior distribution often "mimics" bootstrap histogram of a parameter, including skew. Sort of "smooths" the bootstrap, e.g. logistic regression.
- *Bootstrap Methods and Their Application* by Davison and Hinkley good starting point.