Lecture 16: The Bootstrap

Jacob Steinhardt

March 11, 2021

Recap

- Overdispersion
  - Parametric confidence intervals $\implies$ overly narrow uncertainty
  - Last time: can fix with negative binomial model
  - Are there more model-agnostic ways to fix this?
- Yes! (Sort of)
  - The bootstrap: a nonparametric method for generating confidence intervals
  - Can work even if CLT doesn't hold
  - But can sometimes fail, and need β to at least be meaningful

Recap of frequentist inference

Data $X_1, \ldots, X_n \sim p$, parameter $\theta(p)$

Confidence interval at level α: $I(X_1, \ldots, X_n)$ (interval on real line) such that

$$\mathbb{P}[\theta(p) \in I(X_{1:n})] \geq 1 - \alpha$$

More generally: **confidence region** satisfies $\theta(p) \in R(X_{1:n})$ w.p. $1 - \alpha$.

Note probability is over random draw of $X_1, \ldots, X_n$ (for fixed p).

Wald confidence ellipsoids for GLMs

Last time looked at `statsmodels` package, which uses the Wald ellipsoid:

$$R_\alpha(X_{1:n}) = \{z \mid (z - \hat{\beta}_n)^T \mathrm{I}_n (z - \hat{\beta}_n) \leq F^{-1}(\alpha)\},$$

where $\hat{\beta}_n = \operatorname{argmin}_\beta L_n(\beta)$ is the maximum likelihood estimate, and $\mathrm{I}_n = \nabla^2 L_n(\hat{\beta}_n)$ is the Fisher information.

Asymptotic normality implies that F is the cdf of the χ^2 distribution.

The above form is specific to maximum likelihood estimators, but similar confidence ellipsoids exist for any M-estimator.

Escaping model mis-specification

Saw last time that Wald confidence interval can be wrong if model is wrong

We'll escape this with a **non-parametric** tool for producing frequentist CIs

Non-parametric $\implies$ doesn't rely on model $\implies$ more robust

Key tool: the **bootstrap**

The Bootstrap

Idea for computing confidence intervals by resampling the data

Without bootstrap:

- Often rely on model assumptions
- Wald test, chi-square test, student-t test, ...
- Lots of algebra, need different formula for each setting

With bootstrap:

- Fewer assumptions
- Single unified approach
- Computer simulation

Bootstrap: formal setting

Data: $X^{(1)}, \ldots, X^{(n)} \sim p$

Estimator: $\hat{\theta} = \hat{\theta}(X^{(1)}, \ldots, X^{(n)})$

- θ^*: population parameter (that $\hat{\theta}$ converges to as $n \to \infty$)

Question: How close is θ^* to $\hat{\theta}$?

- Typically framed as computing distribution of $\frac{1}{\sqrt{n}}(\hat{\theta} - \theta^*)$

The ideal hypothetical: re-sampling

Population distribution p^*

- $X^{(1)}, \ldots, X^{(n)} \sim p^*$

The ideal hypothetical: re-sampling

Population distribution p^*

- $X^{(1)}, \ldots, X^{(n)} \sim p^*$

Noise in $\hat{\theta}$ due to randomness in $X^{(1)}, \ldots, X^{(n)}$

The ideal hypothetical: re-sampling

Population distribution p^*

- $X^{(1)}, \dots, X^{(n)} \sim p^*$

Noise in $\hat{\theta}$ due to randomness in $X^{(1)}, \dots, X^{(n)}$

Imagine hypothetically sampling fresh data:

$$
\begin{aligned}
X^{(1)}, \dots, X^{(n)} &\to \hat{\theta} \text{ (Original sample)} \\
X^{(1)\prime}, \dots, X^{(n)\prime} &\to \hat{\theta}' \text{ (Re-sample)} \\
X^{(1)\prime\prime}, \dots, X^{(n)\prime\prime} &\to \hat{\theta}'' \\
X^{(1)\prime\prime\prime}, \dots, X^{(n)\prime\prime\prime} &\to \hat{\theta}''' \\
&\vdots
\end{aligned}
$$

The ideal hypothetical: re-sampling

Population distribution p^*

- $X^{(1)}, \dots, X^{(n)} \sim p^*$

Noise in $\hat{\theta}$ due to randomness in $X^{(1)}, \dots, X^{(n)}$

Imagine hypothetically sampling fresh data:

$$\begin{aligned} X^{(1)}, \dots, X^{(n)} &\to \hat{\theta} \text{ (Original sample)} \\ X^{(1)\prime}, \dots, X^{(n)\prime} &\to \hat{\theta}' \text{ (Re-sample)} \\ X^{(1)\prime\prime}, \dots, X^{(n)\prime\prime} &\to \hat{\theta}'' \\ X^{(1)\prime\prime\prime}, \dots, X^{(n)\prime\prime\prime} &\to \hat{\theta}''' \\ &\vdots \end{aligned}$$

Implicit commitment: distribution of $\hat{\theta}$ roughly centered on θ^* (low bias)

Counterexample

$\hat{\theta}(x_1, \ldots, x_n) = \max_{i=1}^{n} x_i$

n samples: always finite

∞ samples: infinite

The Boostrap

Want to approximate hypothetical samples $\hat{\theta}', \hat{\theta}'', \dots$

But only have actual data $x^{(1)}, \dots, x^{(n)} \to \hat{\theta}$

Idea: subsample data

- With replacement
- n points in each sample

Useful framing: approximate n samples from p by n samples from $\hat{p}_n$

Bootstrap: Pseudocode

B: number of bootstrap samples

For $b = 1, \ldots, B$:

- Sample $x^{(1)\prime}, \ldots, x^{(n)\prime}$ with replacement from $x^{(1)}, \ldots, x^{(n)}$
- Let $\hat{\theta}^{(b)} = \hat{\theta}(x^{(1)\prime}, \ldots, x^{(n)\prime})$

Output $\{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}\}$

Bootstrap in python

[Jupyter demos]

When does the bootstrap work?

Most parametric estimators are fine

- I.e. fixed number of parameters d and $d \ll n$

When does the bootstrap work?

Most parametric estimators are fine

- I.e. fixed number of parameters d and $d \ll n$

NOT parametric:

- Decision trees
- Neural nets
- Kernel regression

These "interpolate" data, sampling with replacement $\approx$ subsampling

When does the bootstrap work?

Most parametric estimators are fine

- I.e. fixed number of parameters d and $d \ll n$

NOT parametric:

- Decision trees
- Neural nets
- Kernel regression

These “interpolate” data, sampling with replacement $\approx$ subsampling

Other commitments:

- $\hat{\theta}$ approximately unbiased
- θ^* is a meaningful quantity

More examples

Bootstrap works for:

- Median and other quantiles
- Cumulative distribution function
- Trimmed mean
- Most U-statistics

Doesn't work for:

- De-generate U-statistics, e.g.: $U(X_{1:n}) = \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbb{I}[X_i = X_j] e^{1/X_i}$
- Estimating θ for $X \sim \mathrm{Uniform}([0, \theta])$.

Bootstrap: Underlying Theory

We seek to approximate the distribution of some quantity $R_n(X_1, \dots, X_n; p)$ for $X_{1:n} \sim p$

Let $\mathscr{L}(p)$ denote the limiting distribution as $n \to \infty$

For instance, $R(X_{1:n}, p) = \frac{\hat{\mu}_n - \mu(p)}{\sqrt{n}\sigma(p)}$, and $\mathscr{L}(p) = N(0, 1)$

Bootstrap replaces $R_n(X_{1:n}, p)$ with $R_n(X'_{1:n}, \hat{p}_n)$

- Intuitively this replaces $\mathscr{L}(p)$ with $\mathscr{L}(\hat{p}_n)$

Bootstrap: Underlying Theory

Bootstrap replaces $R_n(X_{1:n}, p)$ with $R_n(X'_{1:n}, \hat{p}_n)$

- Intuitively this replaces $\mathscr{L}(p)$ with $\mathscr{L}(\hat{p}_n)$

Issue is there are two limits happening at once. To make this work need:

- $R_n(q) \to \mathscr{L}(q)$ uniformly for q in a neighborhood of p
- The mapping $p \mapsto \mathscr{L}(p)$ is continuous

Proof sketch: uniform convergence means that for large n, $R_n(\hat{p}_n)$ will be very close in law to $\mathscr{L}(\hat{p}_n)$ (need uniformity since $\hat{p}_n$ is changing). Then $\mathscr{L}(\hat{p}_n) \to \mathscr{L}(p)$ since $\hat{p}_n \to p$ and $\mathscr{L}$ is continuous.

See Bickel and Freedman 1981, *Some Asymptotic Theory for the Bootstrap*.

Counterexamples Revisited

Nonparametric models (i.e. neural nets) fail because $\mathscr{L}$ is not continuous

Other estimators can fail due to lack of uniformity.

- E.g. $X \sim U([0, \theta])$, take $R_n = \frac{\theta - X^{\max}}{n\theta}$. [Here $X^{\max}$ is the max of the X_i]
- R_n converges to exponential distribution, but bootstrap samples have $X^{\max} = X'^{\max}$ with probability $1 - e^{-1}$.

Some models with growing dimension are actually fine. E.g. can have dimension $n^{1-\delta}$ in regression models and still have bootstrap work. See Mammen 1992, *Bootstrap, wild bootstrap, and asymptotic normality*.