Lecture 12: Robust Uncertainty via the Bootstrap

Jacob Steinhardt

October 7, 2021
Congratulations on finishing Midterm 1!!

Vitamin out today, due Sunday

HW3 out today, due in 2 weeks

Lab and discussion resume as normal next week
Recap

- Bayesian, frequentist regression
- Model mis-specification $\implies$ overly narrow uncertainty (for both Bayesian and frequentist)
- Model checking
Recap

- Bayesian, frequentist regression
- Model mis-specification $\implies$ overly narrow uncertainty (for both Bayesian and frequentist)
- Model checking

This time: more robust frequentist uncertainty estimates, via bootstrap
Bayesian vs. frequentist uncertainty

**Credible interval:** Posterior probability that $\theta$ lies in interval is at least $p$

**Confidence interval:** Conditional on $\theta$, interval contains $\theta$ with probability $p$
Bayesian vs. frequentist uncertainty

**Credible interval:** Posterior probability that $\theta$ lies in interval is at least $p$

**Confidence interval:** Conditional on $\theta$, interval contains $\theta$ with probability $p$

- Another interpretation: no matter what the true parameters are, interval contains them 99% of the time (for $p = 0.99$)
- This property is called *coverage*
Why is a credible interval not (necessarily) a valid confidence interval?

- If true $\theta$ has low prior probability, might not have coverage
Confidence vs. credible intervals

Why is a credible interval not (necessarily) a valid confidence interval?
- If true $\theta$ has low prior probability, might not have coverage

Why is a confidence interval not (necessarily) a valid credible interval?
- Suppose you observe 6 coin flips that all come up heads, but you have very high prior probability that the coin is fair. The 95% confidence interval won’t contain $\frac{1}{2}$, but the credible interval should.
Confidence vs. credible intervals

Other distinction: source of randomness

- Credible interval: randomness is over $\theta$ (posterior probability)
- Confidence interval: randomness is over $X$ (sample of the data)
Confidence vs. credible intervals

Other distinction: source of randomness

- Credible interval: randomness is over $\theta$ (posterior probability)
- Confidence interval: randomness is over $X$ (sample of the data)

Confidence interval requires imagining hypothetical “other” draws of data. We’ll see this used later for the bootstrap.

We’ll focus on confidence intervals for the rest of this lecture.
Confidence intervals for regression

Recall wind turbines example:
\[ \mathbb{E}[\text{Turbines} \mid \text{Year}] = \exp(\beta_{\text{Year}} \cdot \text{Year} + \beta_{\text{Intercept}}) \]

To understand growth rate, care about coefficient \( \beta_{\text{Year}} \)
Confidence intervals for regression

Recall wind turbines example:
$$\mathbb{E}[\text{Turbines} \mid \text{Year}] = \exp(\beta_{\text{Year}} \cdot \text{Year} + \beta_{\text{Intercept}})$$

To understand growth rate, care about coefficient $\beta_{\text{Year}}$

Previously: MCMC sampling gives us posterior distribution (and hence credible interval) for $\beta_{\text{Year}}$:
Confidence intervals for regression

Recall wind turbines example:
\[ \mathbb{E}[\text{Turbines} \mid \text{Year}] = \exp(\beta_{\text{Year}} \cdot \text{Year} + \beta_{\text{Intercept}}) \]

To understand growth rate, care about coefficient \( \beta_{\text{Year}} \)

What about confidence interval? Can’t use prior.
Recall wind turbines example:
\[ E[Turbines \mid Year] = \exp(\beta_{Year} \cdot Year + \beta_{Intercept}) \]

To understand growth rate, care about coefficient \( \beta_{Year} \)

What about confidence interval? Can’t use prior.

General recipe: use generalization of CLT called “asymptotic normality”
Confidence intervals for regression

Recall wind turbines example:
\[ \mathbb{E}[\text{Turbines} \mid \text{Year}] = \exp(\beta_{\text{Year}} \cdot \text{Year} + \beta_{\text{Intercept}}) \]

To understand growth rate, care about coefficient \( \beta_{\text{Year}} \)

What about confidence interval? Can’t use prior.

General recipe: use generalization of CLT called “asymptotic normality”

Beyond scope of this class, but \texttt{statsmodels} package will do it for us!
Confidence intervals with statsmodels

[Jupyter demo]
Escaping model mis-specification

Frequentist confidence intervals can be wrong if model is wrong

- Just like Bayesian case

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

Non-parametric \(\implies\) doesn’t rely on model \(\implies\) more robust
Escaping model mis-specification

Frequentist confidence intervals can be wrong if model is wrong

- Just like Bayesian case

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

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

You’ve seen this before: the bootstrap
The Bootstrap

Idea for computing confidence intervals + uncertainty

Without bootstrap:

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

With bootstrap:

- Single unified approach
- Computer simulation
- Fewer assumptions
Demo: Mean Estimation

[Jupyter demo]
Data: $x_1, \ldots, x_n$

Estimator: $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$

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

Question: How close is $\theta^*$ to $\hat{\theta}$?
Some concrete examples

Mean of 1-dimensional distribution:

- $x_1, \ldots, x_n \in \mathbb{R}$
- $\hat{\theta}(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \ldots + x_n)$

How close is estimate to the true mean?
Some concrete examples

Mean of 1-dimensional distribution:

- $x_1, \ldots, x_n \in \mathbb{R}$
- $\hat{\theta}(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \ldots + x_n)$

How close is estimate to the true mean?

Regression:

- $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$
- $\hat{\beta}((x_1, y_1), \ldots, (x_n, y_n)) = \text{argmin}_\beta \sum_{i=1}^{n} (y_i - \beta^\top x_i)^2$

How close is $\hat{\beta}$ to population parameters $\beta^*$?
More complex examples

Mixture models

Density estimation

Neural nets? (Actually not...)
The ideal hypothetical: re-sampling

Population distribution $p^*$

- $x_1, \ldots, x_n \sim p^*$
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 $p^*$
- $x_1, \ldots, x_n \sim p^*$

Noise in $\hat{\theta}$ due to randomness in $x_1, \ldots, x_n$

Imagine hypothetically sampling fresh data:

- $x_1, \ldots, x_n \rightarrow \hat{\theta}$ (Original sample)
- $x'_1, \ldots, x'_n \rightarrow \hat{\theta}'$ (Re-sample)
- $x''_1, \ldots, x''_n \rightarrow \hat{\theta}''$
- $x'''_1, \ldots, x''''_n \rightarrow \hat{\theta}'''$
- $\vdots$
The ideal hypothetical: re-sampling

Population distribution $p^*$
- $x_1, \ldots, x_n \sim p^*$

Noise in $\hat{\theta}$ due to randomness in $x_1, \ldots, x_n$

Imagine hypothetically sampling fresh data:

$x_1, \ldots, x_n \rightarrow \hat{\theta}$ (Original sample)
$x_1', \ldots, x_n' \rightarrow \hat{\theta}'$ (Re-sample)
$x_1'', \ldots, x_n'' \rightarrow \hat{\theta}''$
$x_1''', \ldots, x_n''' \rightarrow \hat{\theta}'''$

\ldots

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

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

But only have actual data $x_1, \ldots, x_n \rightarrow \hat{\theta}$

Idea: subsample data

- With replacement
- $n$ points in each sample
Bootstrap: Pseudocode

$B$: number of bootstrap samples

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

- Sample $x'_1, \ldots, x'_n$ with replacement from $x_1, \ldots, x_n$
- Let $\hat{\theta}^{(b)} = \hat{\theta}(x'_1, \ldots, x'_n)$

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

[Jupyter demos]
Counterexample: Max

[Jupyter demo]
Counterexample: Max

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

\( n \) samples: always finite

\( \infty \) samples: infinite
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
- $\theta^*$ is a meaningful quantity
Credible intervals vs. confidence intervals
Confidence intervals in statsmodels
Still depend on assumptions!
Bootstrap more robust (and flexible)