Median bias, HulC, and Valid Inference

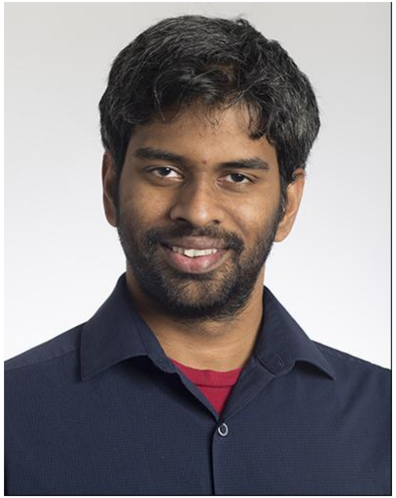
Arun Kumar Kuchibhotla

Carnegie Mellon University

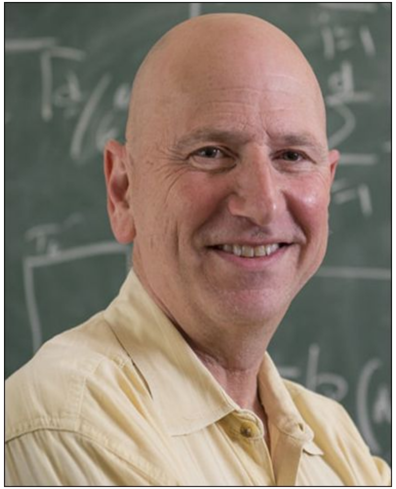
https://arxiv.org/abs/2105.14577
https://arxiv.org/abs/2106.00164

Collaborators

Sivaraman Balakrishnan



Larry Wasserman



Introduction

- Confidence interval is one of the key components of statistical inference.
- Traditional methods of inference are based on the (limiting) distribution of a point estimator.
- There are two “general” methods for construction of confidence intervals:
  - Wald technique: Estimating parameters (e.g., variance) of limiting distribution and using the quantiles of the limiting distribution;
  - Resampling techniques: Estimate the limiting distribution by resampling data and then use quantiles of the estimated distribution.
- Limiting distribution is also a crucial used in defining regularity of an estimator and this in turn is used for discussing uniformly valid inference.

Outline

- Median bias
- HulC
- Coverage/width of HulC intervals
- Simulation Examples
- Valid Inference

Median bias: Introduction

Median bias of an estimator

An estimator $\hat{\theta}_n$ as a function of the data is said to be median unbiased for θ_0 if $\mathrm{Median}(\hat{\theta}_n) = \theta_0$, that is

$$\min\left\{\mathbb{P}(\hat{\theta}_n \leq \theta_0),\, \mathbb{P}(\hat{\theta}_n \geq \theta_0)\right\} \;\geq\; \tfrac{1}{2}.$$

In words, this means that the estimator both over- and under-estimates θ_0 with a probability of at least ½.

In general, we define the median bias of an estimator with respect to a functional as

$$\text{Med-bias}_{\theta_0}(\hat{\theta}_n) \;:=\; \left(\tfrac{1}{2} - \min\left\{\mathbb{P}(\hat{\theta}_n \leq \theta_0),\, \mathbb{P}(\hat{\theta}_n \geq \theta_0)\right\}\right)_+.$$

Median bias: Examples

$$\text{Med-bias}_{\theta_0}(\hat{\theta}_n) := \left(\frac{1}{2} - \min\left\{\mathbb{P}(\hat{\theta}_n \le \theta_0), \mathbb{P}(\hat{\theta}_n \ge \theta_0)\right\}\right)_+.$$

1. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta_0, 1)$, then $\hat{\theta}_n = \overline{X}_n$ is median unbiased. The same holds for any symmetric location family.
2. Suppose $X_1, \ldots, X_n$ are iid with median θ_0, then

   $$\hat{\theta}_n = \begin{cases} X_{(r)}, & \text{with probability } 1/2, \\ X_{(n-r+1)}, & \text{with probability } 1/2, \end{cases}$$

   is median unbiased.
3. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{Unif}(0, \theta_0)$, then $\hat{\theta}_n = 2X_{(n)} - X_{(n-1)}$ is median unbiased.
4. Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$, then $\hat{\theta}_n = \overline{X}_n \mathbf{1}\{\overline{X}_n \ge 0\}$ is median unbiased for $\theta_0 = \mu\mathbf{1}\{\mu \ge 0\}$.

Median bias: Examples (Contd.)

An estimator $\hat{\theta}_n$ is said to be *asymptotically* median unbiased if

$$\lim_{n \to \infty} \text{Med-bias}_{\theta_0}(\hat{\theta}_n) = 0.$$

1. Any estimator with a limiting normal distribution after proper normalization is asymptotically median unbiased.

   This includes asymptotically linear estimators considered in the efficiency framework of parametric and semi-/non-parametric models.

2. Any estimator with a limiting distribution symmetric around zero after proper normalization is asymptotically median unbiased.

   This includes examples from shape constrained literature where the limiting distribution is non-standard, e.g., Chernoff distribution.

Some comments on Median bias

An estimator $\hat{\theta}_n$ is said to be *asymptotically* median unbiased if

$$\lim_{n\to\infty} \text{Med-bias}_{\theta_0}(\hat{\theta}_n) = 0.$$

1. No limiting distribution of the estimator is required to establish its median bias properties.

   E.g.: The sample median is median unbiased for the population median for odd number of observations but sample median need not have a limiting distribution unless the data distribution satisfies some regularity conditions.

2. No moment conditions are required to discuss median bias unlike asymptotic MSE or bias properties.

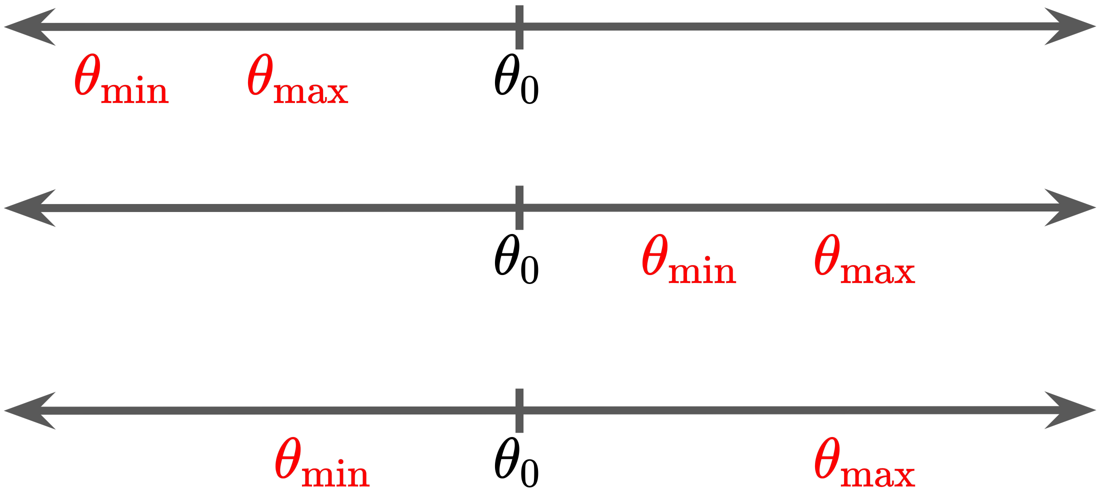
Introducing The HulC

Hull based **C**onfidence

Motivating Calculations

Suppose we have two estimators $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}$, median unbiased for θ_0

$$\theta_{\min} = \min\{\hat{\theta}^{(1)}, \hat{\theta}^{(2)}\}, \quad \text{and} \quad \theta_{\max} = \max\{\hat{\theta}^{(1)}, \hat{\theta}^{(2)}\}.$$



$$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

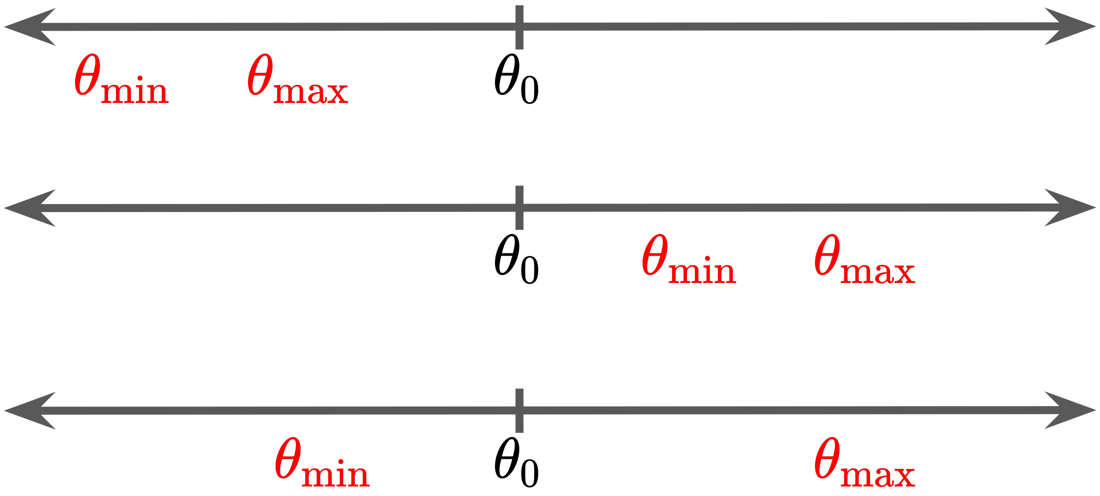
$$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$$\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$$

Simple Calculations

Suppose we have B estimators $\hat{\theta}^{(j)}, 1 \leq j \leq B,$ symmetric about θ_0

$$\theta_{\min} = \min\{\hat{\theta}^{(j)} : j \leq B\}, \quad \text{and} \quad \theta_{\max} = \max\{\hat{\theta}^{(j)} : j \leq B\}.$$



$$\frac{1}{2^B}$$

$$\frac{1}{2^B}$$

$$1 - \frac{2}{2^B}$$

General result

If

Median bias of the estimators.

$$\Delta := \max_{j\geq 1}\left(\tfrac{1}{2} - \min\left\{\mathbb{P}(\hat{\theta}_j \leq \theta_0),\, \mathbb{P}(\hat{\theta}_j \geq \theta_0)\right\}\right)_+,$$

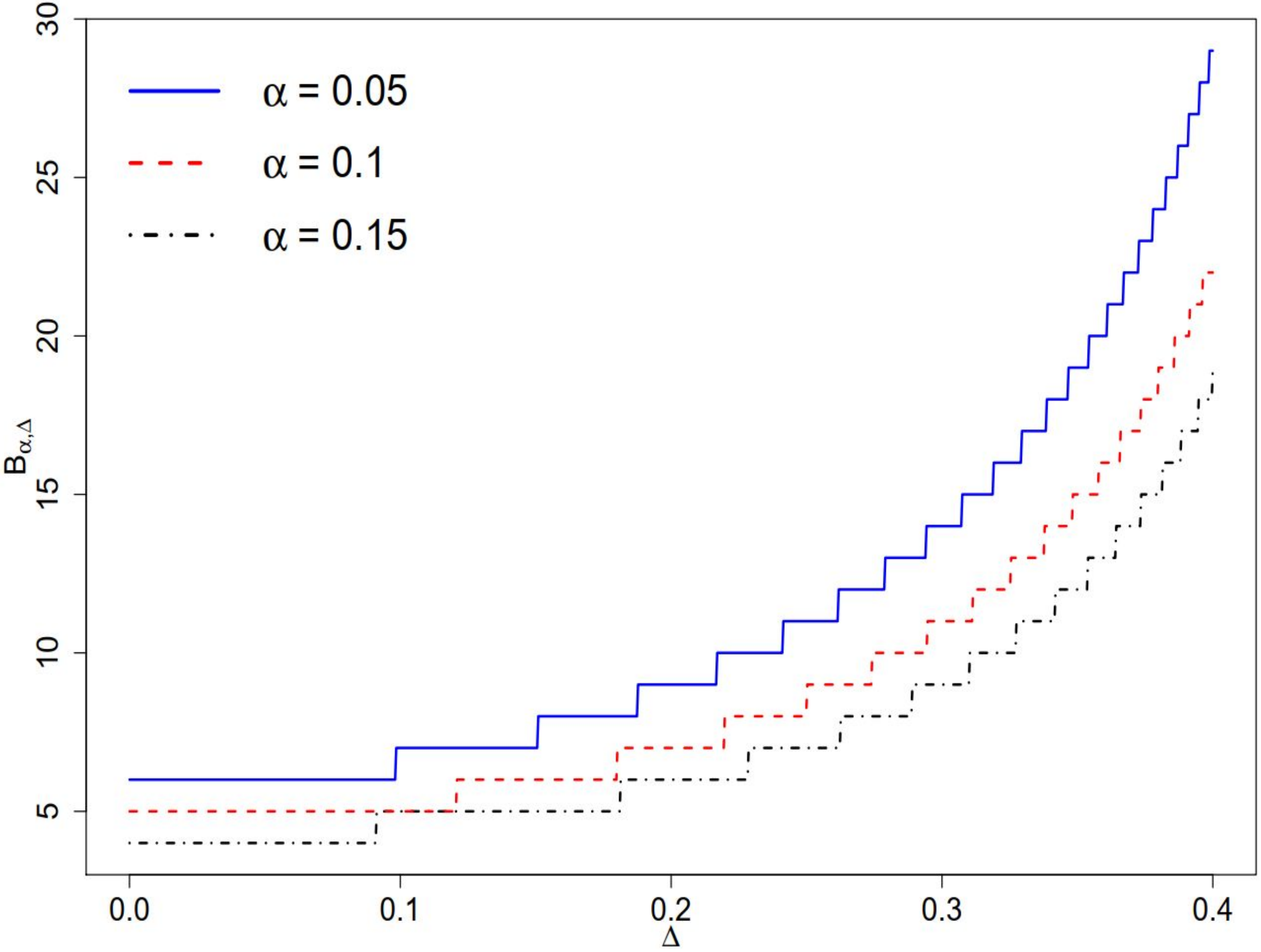
then

$$\begin{aligned}\mathbb{P}\left(\theta_0 \notin [\theta_{\min},\, \theta_{\max}]\right) &\leq P(B;\Delta) \\ &= \left(\frac{1}{2}-\Delta\right)^B + \left(\frac{1}{2}+\Delta\right)^B.\end{aligned}$$

Hence, $B = B_{\alpha,\Delta}$ with $P(B;\Delta)\leq\alpha$ many estimators to get a coverage of at least $1-\alpha.$

Independent copies of the estimator can be obtained by splitting the data.

The number of splits *increase* as the median bias of the estimators *increase*.



Coverage/width of HulC

Coverage of HulC

If

$$\Delta := \max_{j\geq 1}\left(\frac{1}{2} - \min\left\{\mathbb{P}(\hat{\theta}_j \leq \theta_0),\ \mathbb{P}(\hat{\theta}_j \geq \theta_0)\right\}\right)_+,$$

then

$$\begin{aligned}\mathbb{P}\left(\theta_0 \notin \left[\min_{1\leq j\leq B}\hat{\theta}_j,\ \max_{1\leq j\leq B}\hat{\theta}_j\right]\right) &\leq P(B;\Delta) \\ &= \left(\frac{1}{2}-\Delta\right)^B + \left(\frac{1}{2}+\Delta\right)^B.\end{aligned}$$

Take $B = B_{\alpha,0}$ with $P(B;0) \leq \alpha$. A Taylor series expansion yields

0

$$P(B;\Delta) = P(B;0) + P'(B;0)\Delta + O(\Delta^2).$$

Coverage of HulC (cont.)

If

$$\Delta := \max_{j\geq 1}\left(\tfrac{1}{2} - \min\left\{\mathbb{P}(\hat{\theta}_j \leq \theta_0),\ \mathbb{P}(\hat{\theta}_j \geq \theta_0)\right\}\right)_+,$$

and $\Delta = o(1),$ then

$$\mathbb{P}\left(\theta_0 \notin \left[\min_{1\leq j\leq B}\hat{\theta}_j,\ \max_{1\leq j\leq B}\hat{\theta}_j\right]\right) \leq \alpha\left(1 + O(\Delta^2)\right).$$

Hence, the existence of an asymptotically median unbiased estimator implies the existence of an asymptotically valid confidence interval.

Coverage of HulC (cont.)

If

$$\Delta := \max_{j\geq 1}\left(\frac{1}{2} - \min\left\{\mathbb{P}(\hat{\theta}_j \leq \theta_0),\, \mathbb{P}(\hat{\theta}_j \geq \theta_0)\right\}\right)_+,$$

and $\Delta = O(n^{-1/2})$, then

$$\mathbb{P}\left(\theta_0 \notin \left[\min_{1\leq j\leq B}\hat{\theta}_j,\, \max_{1\leq j\leq B}\hat{\theta}_j\right]\right) \;\leq\; \alpha\left(1 + O(n^{-1})\right).$$

Multiplicative error
Second-order accuracy (n^{-1})

Wald and bootstrap confidence intervals have a first-order coverage.
Second-order correct bootstrap intervals exist.

Some notes on coverage

- For asymptotically symmetric estimators, the HulC confidence intervals are second-order accurate.
- This second-order accuracy holds for estimators that do not even converge in distribution.
- Depending only on median bias, HulC operates under weaker conditions than bootstrap and subsampling.
- With estimators with reduced median bias, the HulC confidence intervals are sixth-order accurate (i.e., coverage error of n^{-3}).

Width of HulC Intervals

If the estimators are asymptotically normal with $n^{1/2}$ rate, then the $(1-\alpha)$ confidence HulC and Wald intervals satisfy (asymptotically)

$$\frac{\text{Width of HulC}}{\text{Width of Wald}} = \sqrt{\log_2(\log_2(2/\alpha))} \geq 1.$$

The ratio of widths is larger than one but grows very slowly as $\alpha \to 0.$

In a way, this is the price to pay for the generality of HulC.

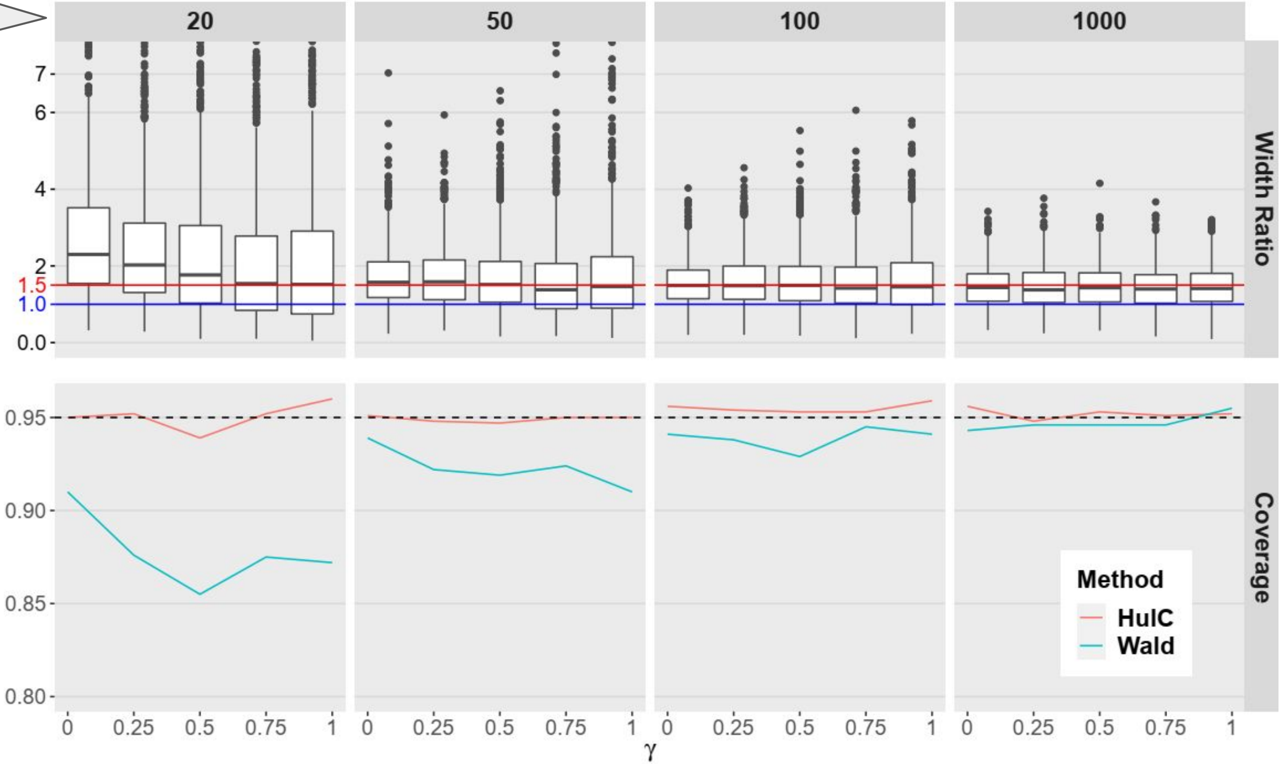
Simulation Examples

OLS Linear Regression

$$Y_i = 1 + 2X_i + \gamma X_i^{1.7} + \exp(\gamma X_i)\xi_i$$

$$X_i \sim \mathrm{Unif}[0, 10], \quad \xi_i \sim N(0, 1).$$

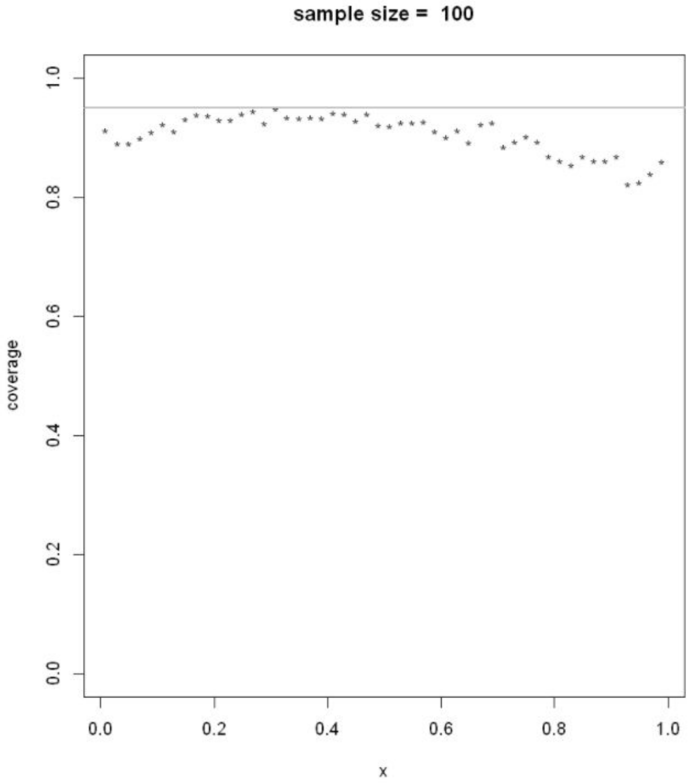
Sample size

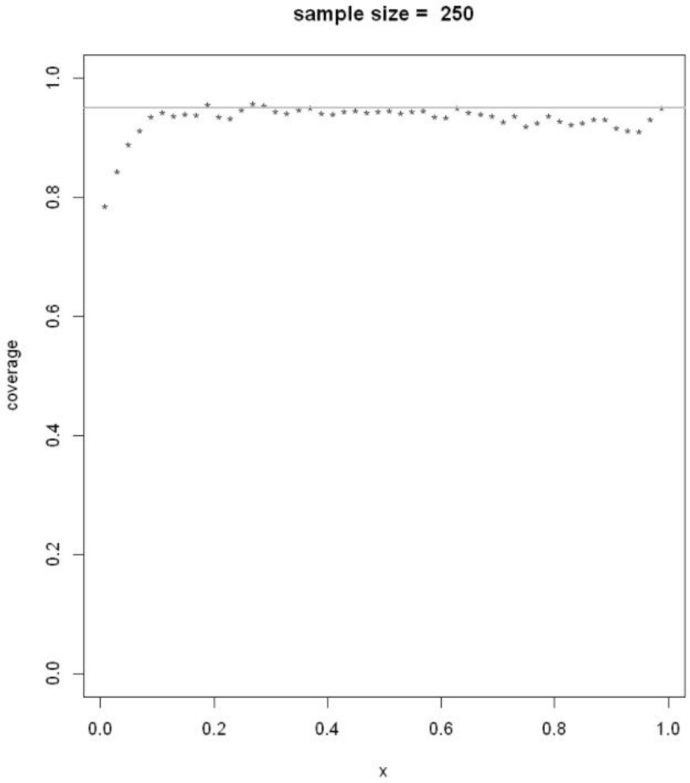


Monotone Regression: LSE

$$Y_i = \exp(2X_i) + \xi_i,$$
$$X_i \sim \text{Unif}[0,1],\ \xi_i | X_i \sim N(0,1)$$

$$\widehat{f} := \underset{f\text{:non-decreasing}}{\text{argmin}} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$





Valid Inference

Coverage of HulC (cont.)

If

$$\Delta := \max_{j\geq 1}\left(\frac{1}{2} - \min\left\{\mathbb{P}(\hat{\theta}_j \leq \theta_0),\, \mathbb{P}(\hat{\theta}_j \geq \theta_0)\right\}\right)_+,$$

and $\Delta = o(1)$, then

$$\mathbb{P}\left(\theta_0 \notin \left[\min_{1\leq j\leq B}\hat{\theta}_j,\ \max_{1\leq j\leq B}\hat{\theta}_j\right]\right) \leq \alpha\left(1 + O(\Delta^2)\right).$$

Hence, the existence of an asymptotically median unbiased estimator implies the existence of an asymptotically valid confidence interval.

Vice versa is also true. Existence of asymptotic valid confidence interval implies the existence of an asymptotically median unbiased estimator.

Inference if and only if HulC

Estimators with a specific limiting median bias is important for applying HulC. Existence of limiting distribution implies this.

Interestingly, if there exists asymptotically valid confidence intervals, then there exists asymptotically median unbiased estimators. If there is a procedure leading to confidence intervals such that for some γ

$$\left(\mathbb{P}(\theta_0 \notin \widehat{\mathbf{CI}}_\gamma) - \gamma\right)_+ \to 0,$$

then there exists an estimator $\hat{\theta}$ such that

$$\text{Med-bias}_{\theta_0}(\hat{\theta}) \to 0.$$

Hence, HulC applies with this estimator and yields a confidence interval with multiplicative miscoverage error and second order accuracy.

Conclusions

- HulC is a general purpose inference method like bootstrap and subsampling.
- Unlike classical methods, HulC does not depend on convergence in distribution but on median bias.
- For asymptotically normal estimators, HulC yields second-order accurate confidence intervals.
- Median bias required for HulC can be estimated via subsampling leading to adaptive HulC.
- Median bias is a crucial ingredient for confidence interval construction.

Thanks for your attention!!