Bootstrapping

Point estimate $\hat{\theta}$. Best guess at θ.

Standard error $SE[\hat{\theta}]$: Summarizes precision of the guess. Formally, is the *estimated* standard deviation of *sampling distribution* of $\hat{\theta}$.

Usual route to getting SE - large-sample theory (e.g., Fisher info.).

Limitations? Computational alternative?

$\bar{Y}$ estimates population mean. $SE = \sqrt{n^{-1}S^2}$.

median$(Y_1, \ldots, Y_n)$ estimates population median. $SE =???$.

Hypothetical/idealized replicated samples

Have *actual* sample of size n from population, giving $\hat{\theta}$.

Draw B further samples (each of size n) from population, yielding $\hat{\theta}_1^{rep}, \ldots, \hat{\theta}_B^{rep}$.

Report SD of $(\hat{\theta}_1^{rep}, \ldots, \hat{\theta}_B^{rep})$ as $SE[\hat{\theta}]$.

Virtual ("bootstrap") replicated samples

Have *actual* sample of size n from population, giving $\hat{\theta}$.

Draw B further samples (each of size n) **by sampling WITH REPLACEMENT from the actual sample**, yielding $\hat{\theta}_1^{rep}, \ldots, \hat{\theta}_B^{rep}$.

Report SD of $(\hat{\theta}_1^{rep}, \ldots, \hat{\theta}_B^{rep})$ as $SE[\hat{\theta}]$.

Example #1.

$Y_1, \ldots, Y_n \overset{iid}{\sim} f$, $n = 100$.

$\hat{\theta} = med(y_1, \ldots, y_n) = 0.58$ estimates $\theta = med(f)$.

Generate $B = 500$ bootstrap samples.
SD of their medians is 0.11.
So report $SE[\hat{\theta}] = 0.11$.

Contrast: asymptotic theory not so easy to apply.

$$\hat{\theta} \overset{approx}{\sim} N\left(\theta, \frac{1}{4n\{f(\theta)\}^2}\right).$$

Need density estimate. Get $SE[\hat{\theta}] = 0.12$.

95% confidence interval. Different possibilities.

1. "Normality-based":

$\hat{\theta} \pm 1.96 SE[\hat{\theta}]$

(0.36, 0.80) in our example.

2. Percentile method:

0.025 and 0.975 percentiles of $\hat{\theta}_1^{rep}, \ldots, \hat{\theta}_B^{rep}$.

(a,b)=(0.46, 0.89) in our example.

3. "Basic" method:

$\left\{ \hat{\theta} - (b - \hat{\theta}), \hat{\theta} + (\hat{\theta} - a) \right\}$

Interpretation: flip percentile interval around $\hat{\theta}$

Justification???

(0.27, 0.70) in our example.

Bootstrapping for more complex data structures

e.g., Data (x_i, y_i), $i = 1, \ldots, n$.

METHOD 1: Generate a bootstrap sample by "re-sampling" (x_i, y_i) pairs.

BUT ... regression models describe $(Y|X)$, not X as well. Suggests...

METHOD 2:

"Fix" $x_1, \ldots, x_n$. Resample residuals $e_i = y_i - \hat{\beta}' x_i$. Add the resampled residuals to the fitted values to generate the Y values in the boostrap sample.

See Example #2: Method 2 does yield smaller SEs than Method 1, as intutions suggests.

Example #3. More complex still.

Simple **model selection** scheme for logistic regression:
1. Fit model with all predictors
2. Remove predictors with $|\hat{\beta}|/SE[\hat{\beta}] < 1.75$.
3. Re-fit with remaining predictors only.

SEs from final fit do not reflect uncertainty about which predictors to keep/discard.

Can a bootstrap SE fix this problem?

Example suggest yes.

NOTE: lots of uncertainty about how many and which predictors to keep. *Model selection is unstable!*

Focus on the PTD predictor. SE from final fit is 0.44.

Bootstrap SE (conditional on inclusion of PTD) is 0.51.

Bootstrap SE (unconditionally) is 0.78.