# Introduction to the Bootstrap

Methods Seminar VIII

Charles Thomas

clt6@po.cwru.edu

## Summary

- Will describe
  - basic ideas
  - confidence intervals
  - application to hypothesis testing and regression problems
  - examples

## Bootstrap defined

- The bootstrap is a form of a larger class of methods that resample from the original data set and therefore are called resampling procedures
- Some resampling procedures go back a long way (e.g. the jackknife—1949, permutation methods—1930s)
- Computer based method for assigning measures of accuracy to statistical estimates(Efron, 1998)
- Efron along with colleagues connected the nonparametric bootstrap(resampling with replacement) with earlier accepted statistical tools such as the jackknife and delta method for estimating standard errors

## Bootstrap process described

- B bootstrap samples are generated from the original dataset
- Each bootstrap sample has **n** elements, generated by sampling with replacement *n* times
- Bootstrap replicates $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \ldots\ldots, s(\mathbf{x}^{*B})$ are obtained by calculating the value of the estimator of the replicate
- The standard deviation of the values $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \ldots\ldots, s(\mathbf{x}^{*B})$ is the estimate of standard error of $s(\mathbf{x})$, sometimes called the Monte Carlo approximation to the bootstrap estimate of the standard error

## Bootstrap process described(cont)

- We really would like to know the distribution of $\hat{s}(x)$– s(**x**)
- What we have, however, is the Monte Carlo approximation to the distribution of $s(\mathbf{x})^* - \hat{s}(x)$
- With a sufficiently large **n** the two distributions are expected to be nearly the same
- That the distribution of $s(\mathbf{x})^* - \hat{s}(x)$ behaves almost like the distribution of $\hat{s}(x)$ - s(**x**)

## The empirical distribution and the plug-in principle

- Statistical inference involves estimating some aspect of a probability distribution **F**
- A sensible way of estimating some aspect of **F**, such as the mean, median or correlation, is to use the corresponding aspect of **F**
- This is called the plug-in principle
- The bootstrap method is a direct application of the plug-in principle

## The empirical distribution (continued)

- Observe **F** → **($x_1$, $x_2$,…, $x_n$,)**, the empirical distribution **F** is the discrete distribution that puts probability 1/*n* on each value $x_i$, $i$ =1,2, …, *n*.
- **F** assigns to a set **A** in the sample space of *x* its empirical probability **Prob{A} = #{$x_i$ Є A}/*n***
- The probability is really the proportion of occurrence of each value in the empirical distribution
- Information is not lost going from the full data set (sample space) to the reduced

## The Empirical distribution (cont)

- It is true that the vector of observed frequencies $\hat{F} = (\hat{f}_1, \hat{f}_2, \ldots)$ is a sufficient statistic for the distribution **F** = $(f_1, f_2, \ldots)$
- All the information about **F** contained in **x** is also contained in **F**
- The sufficiency theorem assumes that the data have been generated by random sampling from some distribution **F**
- This is not always true

## The plug-in principle

- Simple method of estimating parameters from samples
- The plug-in estimate of a parameter $\theta = t(\mathbf{F})$ is defined to be $\hat{\theta} = t(\hat{F})$
- The bootstrap is used to study the bias and standard error of the plug-in estimate
- The bootstrap produces biases and standard errors in an automatic fashion
- The plug-in principle is less good when there is information about **F** other than that provided by the sample **x**
- The plug-in principle and the bootstrap can be adapted to parametric families and regression models

## Statistics and standard errors

- Summary statistics are often the first outputs of a data analysis
- The bootstrap provides accuracy estimates by using the plug-in principle
- The bootstrap estimate of standard error requires no theoretical calculations
- It is available even if the estimator $\theta = s(\mathbf{x})$ is mathematically complex

## The bootstrap estimate of standard error

- Bootstrap methods for estimating standard errors depend upon the bootstrap sample
- Corresponding to a bootstrap data set $\mathbf{x}^*$ is a bootstrap replication of $\theta = s(\mathbf{x}^*)$
- The bootstrap estimate the standard error of a statistic is a plug-in estimate that uses the empirical distribution function $\hat{F}$ in place of the unknown distribution **F**

## The bootstrap algorithm for estimating standard errors

1. Select B independent bootstrap samples $\mathbf{x}^{*1}\ \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$, each consisting of $n$ data values drawn with replacement from **x**

2. Evaluate the bootstrap replication corresponding to each bootstrap sample,
   $\hat{\theta}(b) = s(\mathbf{x}^{*b}) \quad b = 1,2, \ldots, B$

## The bootstrap algorithm for estimating SE(cont)

3. Estimate the standard error by the sample standard deviation of the $B$ replications

$$\hat{se}_B = \{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2/(B-1)\}^{1/2},$$

Where $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B}\hat{\theta}^*(b)/B, b = 1,2,\ldots,B$

## The number of bootstrap replications *B*

- A small number of replications $B$ = 25, is usually informative according to Efron
- 50 replications is often enough to give good estimate of standard error
- Much bigger values of $B$ are required for bootstrap confidence intervals
- 1000 replications is recommended by Harrell and others for stable confidence intervals

## The parametric bootstrap

- Bootstrap resampling carried out parametrically
- Standard error from this process will closely resemble results derived from textbook formulae
- Why conduct bootstrap process if theory and formulae have been developed?
- Where is bootstrap process inferior to formula application
- Davison and Hinkley (1997) justify the nonparametric bootstrap in parametric problems as a test of robustness of validity of the parametric method

## Estimation of bias

- Another measure of statistical accuracy is bias
- Bias is the difference between the expected value of an estimator and the quantity being estimated
- The bootstrap estimate of bias is:
  $bias_{\hat{F}} = E_{\hat{F}}[s(x^*)] - t(\hat{F}), where\ t(\hat{F})\ is\ the\ plug-in\ estimate\ of\ \theta$

## Estimation of bias(cont)

- If s(**x**) is the mean it can be shown that $bias_{\hat{F}} = 0$
- Estimates of bootstrap bias must be done with Monte Carlo simulation with B replications of the form

$$bias_B = \hat{\theta}^*(\cdot) - t(\hat{F}), \ \ where \ \hat{\theta}^*(\cdot) = \sum_{b=1}^{B} s(x^{*b})/B$$

## Confidence intervals

- Standard errors are often used to assign approximate confidence intervals to a parameter of interest
- This parameter is assumed to be normally distributed with known standard error
- The random quantity $Z = (\hat{\theta} - \theta)/\hat{s}e \sim N(0,1)$

valid for $n \rightarrow \infty$

but a limited approximation for finite samples

## Confidence intervals

- The standard confidence interval can be improved upon using the t distribution for finite samples
- The use of the t distribution does not adjust for the skewness of the underlying distribution or other errors that result when $\hat{\theta}$ is not the sample mean

## Confidence intervals based on the bootstrap-t interval

- Obtain accurate intervals without having to make normal theory assumptions
- Estimate the distribution of Z directly from the data
- Can build a table of values for this process as you can for the Normal and t distributions
- The bootstrap table is built by generating B bootstrap samples and computing the bootstrap version of Z for each

$$Z^*(b) = (\hat{\theta}^*(b) - \hat{\theta})/\hat{s}e^*(b), \ where \ \hat{\theta}^*(b) = s(x^{*b})$$

## Table of percentiles

| Percentile | 5% | 10% | 16% | 50% | 84% | 90% | 95% |
|---|---|---|---|---|---|---|---|
| $t_5$ | -2.01 | -1.48 | -1.73 | 0 | 1.73 | 1.48 | 2.01 |
| $t_8$ | -1.86 | -1.40 | -1.10 | 0 | 1.10 | 1.40 | 1.86 |
| $t_{20}$ | -1.73 | -1.33 | -1.06 | 0 | 1.06 | 1.33 | 1.73 |
| $t_{50}$ | -1.68 | -1.30 | -1.02 | 0 | 1.02 | 1.30 | 1.68 |
| $t_{100}$ | -1.66 | -1.29 | -1.00 | 0 | 1.00 | 1.29 | 1.66 |
| Normal | -1.65 | -1.28 | -0.99 | 0 | 0.99 | 1.28 | 1.65 |
| Bootstrap-t | -4.53 | -2.01 | -1.32 | -.025 | 0.86 | 1.19 | 1.53 |

## Confidence intervals based on bootstrap percentiles

- Percentiles of the bootstrap histogram define the confidence limits
- If the bootstrap distribution of $\hat{\theta}$ is roughly normal, then the standard normal and percentile intervals will nearly agree
- The percentile method automatically makes transformation if such transformation exists

## Bias corrected confidence intervals $BC_a$

- Bootstrap intervals should match exact confidence intervals where statistical theory provides an exact answer
- These intervals should also give dependably good coverage properties in all situations
- Neither bootstrap-t method nor the percentile method meet both of the above criteria
- $BC_a$, a version of the percentile method, corrects the problems with the other methods

## Regression models and the bootstrap

- The general regression model:

  $$Y_i = g_i(\beta) + \varepsilon_i \ \ for\ i = 1,2,\ldots,n$$

- g is of known form and may depend on a fixed vector of covariates, $\beta$ is a vector of unknown parameters and  are independently and identically distributed with some distribution **F**

## Regression and the bootstrap (cont)

- Denote distance measure $D(y,\lambda(\beta)) = \sum_{i=1}^{n}[y_i - g_i(\beta)]^2$ we get least squares estimates
- $\hat{\beta} = \min(D(y,\lambda(\beta)))$
- The residuals are obtained by $\hat{\varepsilon}_i = y_i - g_i(\beta)$
- The first bootstrap approach is to bootstrap the residuals $\varepsilon_i$

## Regression and the bootstrap (cont)

- Construct a bootstrap sample data set $y_i^* = g(\hat{\beta}) + \varepsilon_i^*,\ for\, i = 1,2,\ldots,n$
- Sample $\varepsilon_i^*$ with replacement B times
- Calculate $\hat{\beta}^* = \min(D(y,\lambda(\beta)))$

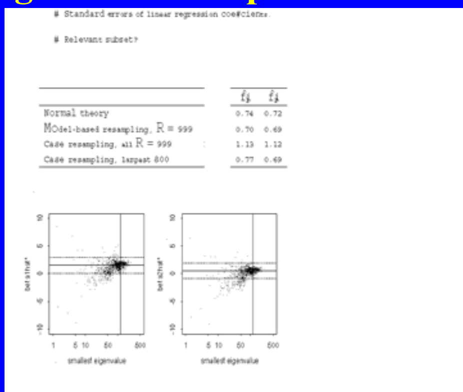
## Regression and the bootstrap (cont)

- A second approach is to bootstrap $z_i=(y_i,c_i)$ of the observations $y_i$ and covariates $c_i$
- The bootstrap samples are then $z^*= (y^*,c^*)$
- $y^*$ is used to obtain the estimate $\hat{\beta}^*$ just as before
- This method is less sensitive to modeling assumptions

## Regression example

- From Anthony Davison, 1999. Simulated data consisting of 13 observations
- Dependent variable continuous from relatively uniform distribution
- Covar compositional $x_1 + \ldots\ldots + x_4 \approx 100$ so **X** almost collinear

## Regression example


# Standard errors of linear regression coefficients.

# Relevant subset?

| | [illegible] | [illegible] |
|---|---|---|
| Normal theory | 0.74 | 0.72 |
| Model-based resampling, R = 999 | 0.70 | 0.69 |
| Case resampling, all R = 999 | 1.13 | 1.12 |
| Case resampling, largest 800 | 0.77 | 0.69 |



## Why not just use least squares?

- Least squares estimates are very sensitive to violations of the modeling assumptions
- If error distribution is not Gaussian, the bootstrap provides a method for computing standard errors or prediction intervals regardless of method of estimation
- Other complications to the regression problem such as heteroscedasticity and nonlinearity in the model terms

## Hypothesis testing and the bootstrap – two sample

- $H_o$: $F=G$
- Test statistic is denoted by $t(x) = \bar{z} - \bar{y}$ the difference in means(need not be an estimate of a parameter)
- Computation of bootstrap test stat
  1. Draw B samples of size n+m with replacement
  2. Evaluate t(·) on each sample, $t(x^{*b}) = \bar{z}^* - \bar{y}^*$
  3. Approx. $ASL_{boot}$ by

$\hat{ASL}_{boot} = \#\{t(x^{*b}) \geq t_{obs\}} / B,$

$t_{obs} = t(x)$ *the oberved value of the statistic*

## Hypothesis testing and the bootstrap – one sample

- A one sample version of the normal test could be used
- Assume normality under $H_{o,}$
- ASL = $\phi(\bar{z}^* - t)/(\sigma/\sqrt{n})$ where $\phi$ is the cumulative distribution function of the standard normal

## Estimates of location and dispersion and the bootstrap

- First and second moments of known distributions allow for the calculation of population parameters
- The population mean is the natural location parameter
- Distributions without first moments, the median is a natural location parameter
- Bootstrapping is useful, not in point estimation, but in providing measures of dispersion and measures of accuracy

## Estimates of location and dispersion (cont)

- For distributions whose moments are undefined(e.g. the Cauchy distribution), the sample mean does not converge to the population mean
- The median however, does converge
- If nothing is known about the population, then estimating the median is probably the correct approach

## In contrast to the jackknife

- Originally goal was to improve an estimate of bias
- Became more useful as a way to estimate variances and standard errors
- Focuses on the samples that leaves out one observation at a time
- Although predates the bootstrap, it bears strong resemblance to the bootstrap
- The bootstrap is generally considered to be more efficient

## In contrast to permutation methods

- Based on order statistic representation, meaning that all possible permutations of the data vector are chosen and analyzed
- Samples are not drawn with replacement
- The bootstrap gives very similar results to the permutation test

## Example

- Adapted from Fox (1997) "Applied Regression Analysis"
- Goal: Estimate mean difference between Male and Female finding X
- Four pairs of observations are available:

## Tables of values

| Observ. | Male | Female | Differ. Y |
|---|---|---|---|
| 1 | 24 | 18 | 6 |
| 2 | 14 | 17 | -3 |
| 3 | 40 | 35 | 5 |
| 4 | 44 | 41 | 3 |

## Mean Difference

Sample mean is 2.75

- If Y were normally distributed, 95% CI

$$\mu = \overline{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- But we do not know $\sigma$

## Estimates

- Estimate of $\sigma$ is $s = \sqrt{\frac{\sum (Y_j - \overline{Y})^2}{(n-1)}}$
- Estimate of standard error is $S\hat{E}(\overline{Y}) = \frac{S}{\sqrt{n}} = 2.015$
- Assuming population is normally distributed, we can use t-distribution as

$$\mu = \overline{Y} \pm t_{n-1,0.025} \frac{S}{\sqrt{n}}$$

## Confidence Intervals

Very Wide

$$-5.91 < \mu < 11.41$$

## Bootstrap sample mean, variance and SE

- Use distribution Y* of sample to estimate distribution Y in population
- After sampling with replacement, B=1000, we get:
  E*(Y*)=2.74
  SE of the bootstrap replicates is 1.76
- This is smaller than the SE calculated from the sample

## Results of bootstrapping

Observed Bias Mean SE

- Param 2.75 -0.01475 2.735 1.763
- Empirical Percentiles:

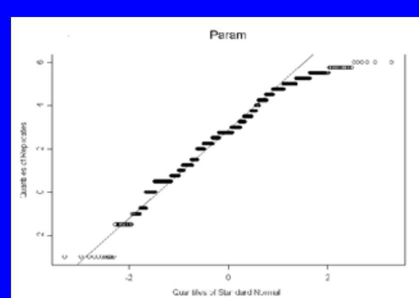
  2.5% 5% 95% 97.5%

- Param -1.5 0 5.25 5.5
- BCa Percentiles:

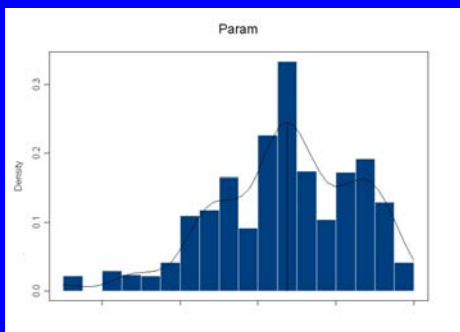
  2.5% 5% 95% 97.5%

- Param -2.205 -1.5 4.75 5

## Normal qq-plot of replicated mean differences



# Distribution of bootstrap replicates




# References

Chernick, M. (1999). *Bootstrap Methods: A Practitioner's Guide*. John Wiley & Sons, New York.

Efron, B, and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Manly B. (1997). *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. Chapman & Hall, New York.

Efron, B. (1979a). Bootstrap methods: another look at the jackknife. Ann. Statist. 7, 1-26.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? J. Am. Statist. Assoc. 81, 461-470.

Efron, B. (1987). Better bootstrap confidence intervals. J. Am. Statist. Assoc. 82, 171-200.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors: confidence intervals and other measures of statistical accuracy. Statist. Sci. 1, 54-77.