Bootstrap, Jackknife and other resampling methods

Part II: Non-Parametric Bootstrap

Rozenn Dahyot

Room 128, Department of Statistics
Trinity College Dublin, Ireland
dahyot@mee.tcd.ie

2005

Introduction

We want to assess the accuracy (bias, standard error, etc.) of an arbitrary estimate $\hat{\theta}$ knowing only one sample $\mathbf{x} = (x_1, \cdots, x_n)$ drawn from an unknown population density function F.

- We propose here one way, called *Bootstrap*, to do it using computer intensive techniques for resampling.
- Bootstrap is a data based simulation method for statistical inference. The basic idea of bootstrap is to use the sample data to compute a statistic and to estimate its sampling distribution, without any model assumption.
- No theoretical calculations of standard errors needed so we don't care how mathematically complex the estimator $\hat{\theta}$ can be!

Introduction

- The (non-parametric) bootstrap method is an application of the plug-in principle. By *non-parametric*, we mean that only **x** is known (observed) and no prior knowledge on the population density function F is available.
- Originally, the Bootstrap was introduced to compute standard error of an arbitrary estimator by Efron (1979) and to-date the basic idea remains the same.
- The term bootstrap derives from the phrase *to pull oneself up by one's bootstrap* (Adventures of Baron Munchausen, by Rudolph Erich Raspe). The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.

Bootstrap samples and replications

Definition

A bootstrap sample $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$ is obtained by randomly sampling n times, with replacement, from the original data points $\mathbf{x} = (x_1, x_2, \cdots, x_n)$.

Considering a sample $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$, some bootstrap samples can be:

$$
\begin{aligned}
\mathbf{x}^{*(1)} &= (x_2, x_3, x_5, x_4, x_5) \\
\mathbf{x}^{*(2)} &= (x_1, x_3, x_1, x_4, x_5) \\
&\text{etc.}
\end{aligned}
$$

Definition

With each bootstrap sample $\mathbf{x}^{*(1)}$ to $\mathbf{x}^{*(B)}$, we can compute a bootstrap replication $\hat{\theta}^*(b) = s(\mathbf{x}^{*(b)})$ using the plug-in principle.

How to compute Bootstrap samples

Repeat B times:

1. A random number device selects integers $i_1, \cdots, i_n$ each of which equals any value between 1 and n with probability $\frac{1}{n}$.
2. Then compute $\mathbf{x}^* = (x_{i_1}, \cdots, x_{i_n})$.

Some matlab code available on the web

See BOOTSTRAP MATLAB TOOLBOX, by Abdelhak M. Zoubir and D. Robert Iskander,
http://www.csp.curtin.edu.au/downloads/bootstrap_toolbox.html

How many values are left out of a bootstrap resample ?

Given a sample $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and assuming that all x_i are different, the probability that a particular value x_i is left out of a resample $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$ is:

$$\mathcal{P}(x_j^* \neq x_i, 1 \leq j \leq n) = \left(1 - \frac{1}{n}\right)^n$$

since $\mathcal{P}(x_j^* = x_i) = \frac{1}{n}$. When n is large, the probability $\left(1 - \frac{1}{n}\right)^n$ converges to $e^{-1} \approx 0.37$.

The Bootstrap algorithm for Estimating standard errors

1. Select B independent bootstrap samples $\mathbf{x}^{*(1)}, \mathbf{x}^{*(2)}, \cdots, \mathbf{x}^{*(B)}$ drawn from $\mathbf{x}$
2. Evaluate the bootstrap replications:

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*(b)}), \quad \forall b \in \{1, \cdots, B\}$$

3. Estimate the standard error $\mathrm{se}_F(\hat{\theta})$ by the standard deviation of the B replications:

$$\hat{\mathrm{se}}_B = \left[\frac{\sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1}\right]^{\frac{1}{2}}$$

where $\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B}$

Bootstrap estimate of the standard Error

Example A

From the distribution F: $F(x) = 0.2\ \mathcal{N}(\mu{=}1,\sigma{=}2) + 0.8\ \mathcal{N}(\mu{=}6,\sigma{=}1)$. We draw the sample $\mathbf{x} = (x_1, \cdots, x_{100})$:

$$\mathbf{x} = \left\{ \begin{array}{ccccc}
7.0411 & 4.8397 & 5.3156 & 6.7719 & 7.0616 \\
5.2546 & 7.3937 & 4.3376 & 4.4010 & 5.1724 \\
7.4199 & 5.3677 & 6.7028 & 6.2003 & 7.5707 \\
4.1230 & 3.8914 & 5.2323 & 5.5942 & 7.1479 \\
3.6790 & 0.3509 & 1.4197 & 1.7585 & 2.4476 \\
-3.8635 & 2.5731 & -0.7367 & 0.5627 & 1.6379 \\
-0.1864 & 2.7004 & 2.1487 & 2.3513 & 1.4833 \\
-1.0138 & 4.9794 & 0.1518 & 2.8683 & 1.6269 \\
6.9523 & 5.3073 & 4.7191 & 5.4374 & 4.6108 \\
6.5975 & 6.3495 & 7.2762 & 5.9453 & 4.6993 \\
6.1559 & 5.8950 & 5.7591 & 5.2173 & 4.9980 \\
4.5010 & 4.7860 & 5.4382 & 4.8893 & 7.2940 \\
5.5741 & 5.5139 & 5.8869 & 7.2756 & 5.8449 \\
6.6439 & 4.5224 & 5.5028 & 4.5672 & 5.8718 \\
6.0919 & 7.1912 & 6.4181 & 7.2248 & 8.4153 \\
7.3199 & 5.1305 & 6.8719 & 5.2686 & 5.8055 \\
5.3602 & 6.4120 & 6.0721 & 5.2740 & 7.2329 \\
7.0912 & 7.0766 & 5.9750 & 6.6091 & 7.2135 \\
4.9585 & 5.9042 & 5.9273 & 6.5762 & 5.3702 \\
4.7654 & 6.4668 & 6.1983 & 4.3450 & 5.3261
\end{array} \right\}$$

We have $\mu_F = 5$ and $\overline{x} = 4.9970$.

Bootstrap estimate of the standard Error

Example A

1. $B = 1000$ bootstrap samples $\{\mathbf{x}^{*(b)}\}$
2. $B = 1000$ replications $\{\overline{x}^*(b)\}$
3. Bootstrap estimate of the standard error:

$$\widehat{\mathrm{se}}_{B=1000} = \left[\frac{\sum_{b=1}^{1000}[\overline{x}^*(b) - \overline{x}^*(\cdot)]^2}{1000-1}\right]^{\frac{1}{2}} = 0.2212$$

where $\overline{x}^*(\cdot) = 5.0007$. This is to compare with $\hat{se}(\overline{x}) = \frac{\hat{\sigma}}{\sqrt{n}} = 0.22$.

Distribution of $\hat{\theta}$

When enough bootstrap resamples have been generated, not only the standard error but any aspect of the distribution of the estimator $\hat{\theta} = t(\hat{F})$ could be estimated. One can draw a histogram of the distribution of $\hat{\theta}$ by using the observed $\hat{\theta}^*(b)$, $b = 1, \cdots, B$.

Example A

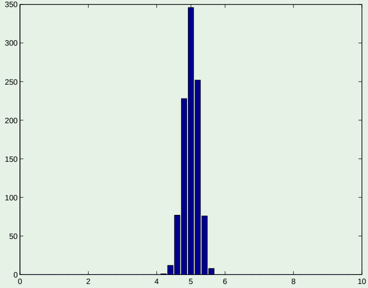

Figure: Histogram of the replications $\{\overline{x}^*(b)\}_{b=1\cdots B}$.

Bootstrap estimate of the standard error

Definition

The ideal bootstrap estimate $\mathrm{se}_{\hat{F}}(\theta^*)$ is defined as:

$$\lim_{B\to\infty} \hat{\mathrm{se}}_B = \mathrm{se}_{\hat{F}}(\theta^*)$$

$\mathrm{se}_{\hat{F}}(\theta^*)$ is called a non-parametric bootstrap estimate of the standard error.

Bootstrap estimate of the standard Error

How many B in practice ?

you may want to limit the computation time. In practice, you get a good estimation of the standard error for B in between 50 and 200.

Example A

B	10	20	50	100	500	1000	10000
$\widehat{\mathrm{se}}_B$	0.1386	0.2188	0.2245	0.2142	0.2248	0.2212	0.2187

Table: Bootstrap standard error w.r.t. the number B of bootstrap samples.

Bootstrap estimate of bias

Definition

The bootstrap estimate of bias is defined to be the estimate:

$$\begin{aligned}\mathrm{Bias}_{\hat{F}}(\hat{\theta}) &= \mathbb{E}_{\hat{F}}[s(\mathbf{x}^*)] - t(\hat{F}) \\ &= \theta^*(\cdot) - \hat{\theta}\end{aligned}$$

Example A

B	10	20	50	100	500	1000	10000
$\mathbb{E}_{\hat{F}}(\overline{x}^*)$	5.0587	4.9551	5.0244	4.9883	4.9945	5.0035	4.9996
$\widehat{\mathrm{Bias}}$	0.0617	-0.0419	0.0274	-0.0087	-0.0025	0.0064	0.0025

Table: $\widehat{\mathrm{Bias}}$ of $\overline{x}^*$ ($\overline{x} = 4.997$ and $\mu_F = 5$).

Bootstrap estimate of bias

1. B independent bootstrap samples $\mathbf{x}^{*(1)}, \mathbf{x}^{*(2)}, \cdots, \mathbf{x}^{*(B)}$ drawn from $\mathbf{x}$
2. Evaluate the bootstrap replications:

$$\hat{\theta}^*(b) = s(\mathbf{x}^{*(b)}), \quad \forall b \in \{1, \cdots, B\}$$

3. Approximate the bootstrap expectation :

$$\hat{\theta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b) = \frac{1}{B} \sum_{b=1}^{B} s(\mathbf{x}^{*(b)})$$

4. the bootstrap estimate of bias based on B replications is:

$$\widehat{\mathrm{Bias}}_B = \hat{\theta}^*(\cdot) - \hat{\theta}$$

Confidence interval

Definition

Using the bootstrap estimation of the standard error, the $100(1-2\alpha)\%$ confidence interval is:

$$\theta = \hat{\theta} \pm z^{(1-\alpha)} \cdot \widehat{\mathrm{se}}_B$$

Definition

If the bias in not null, the bias corrected confidence interval is defined by:

$$\theta = (\hat{\theta} - \widehat{\mathrm{Bias}}_B) \pm z^{(1-\alpha)} \cdot \widehat{\mathrm{se}}_B$$

Can the bootstrap answer other questions?

The mouse data

Data (Treatment group)	94; 197; 16; 38; 99; 141; 23
Data (Control group)	52; 104; 146; 10; 51; 30; 40; 27; 46

Table: The mouse data [Efron]. 16 mice divided assigned to a treatment group (7) or a control group (9). Survival in days following a test surgery. Did the treatment prolong survival ?

Can the bootstrap answer other questions?

The mouse data

- Remember in the first lecture, we compute $d = \overline{x}_{Treat} - \overline{x}_{Cont} = 30.63$ with a standard error $\hat{se}(d) = 28.93$. The ratio was $d/\hat{se}(d) = 1.05$ (an insignificant result as measuring $d = 0$ is likely possible).
- Using bootstrap method
  1. B bootstrap samples $\mathbf{x}^{*(b)}_{Treat} = (x^{*(b)}_{Treat\ 1}, \cdots, x^{*(b)}_{Treat\ 7})$ and $\mathbf{x}^{*(b)}_{Cont} = (x^{*(b)}_{Cont\ 1}, \cdots, x^{*(b)}_{Cont\ 9})$, $\forall 1 \leq b \leq B$
  2. B bootstrap replications are computed: $d^*(b) = \overline{x}^{*(b)}_{Treat} - \overline{x}^{*(b)}_{Cont}$
  3. The bootstrap standard error is computed for $B = 1400$: $\hat{se}_{B=1400} = 26.85$.
  4. The ratio is $d/\hat{se}_{1400}(d) = 1.14$.
- This is still not a significant result.

The Law school example

School	1	2	3	4	5	6	7	8
LSAT (X)	576	635	558	578	666	580	555	661
GPA (Y)	3.39	3.30	2.81	3.03	3.44	3.07	3.00	3.43

School	9	10	11	12	13	14	15
LSAT (X)	651	605	653	575	545	572	594
GPA (Y)	3.36	3.13	3.12	2.74	2.76	2.88	2.96

Table: Results of law schools admission practice for the LSAT and GPA tests. It is believed that these scores are highly correlated. Compute the correlation and its standard error.

Correlation

The correlation is defined :

$$\mathrm{corr}(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))]}{\left(\mathbb{E}[(X - \mathbb{E}(X))^2] \cdot \mathbb{E}[(Y - \mathbb{E}(Y))^2]\right)^{1/2}}$$

Its typical estimator is:

$$\widehat{\mathrm{corr}}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i \ y_i - n \ \overline{x} \ \overline{y}}{[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2]^{1/2} \cdot [\sum_{i=1}^{n} y_i^2 - n\overline{y}^2]^{1/2}}$$

The Law school example

- The estimated correlation is $\widehat{\text{corr}}(\mathbf{x}, \mathbf{y}) = .7764$ between LSAT and GPA.
- Precise theoretical formula for the standard error of the estimator is unavailable.

Non-parametric Bootstrap estimate of the standard error

B	25	50	100	200	400	800	1600	3200
$\hat{\text{se}}_B$	.140	.142	.151	.143	.141	.137	.133	.132

Table: Bootstrap estimate of standard error for $\widehat{\text{corr}}(\mathbf{x}, \mathbf{y}) = .776$.

The standard error stabilizes to $\text{se}_{\hat{F}}(\widehat{\text{corr}}) \approx .132$.

The Law school example: Conclusion

- The textbook formula for the correlation coefficient is:

$$\hat{\mathrm{se}}(\widehat{\mathrm{corr}}) = (1 - \widehat{\mathrm{corr}}^2)/\sqrt{n-3}$$

- With $\widehat{\mathrm{corr}} = 0.7764$, the standard error is $\hat{se}(\widehat{\mathrm{corr}}) = 0.1147$.
- The estimated non-parametric bootstrap standard error $\mathrm{se}_{B=3200}$ is 0.132.

Summary

- Re-sampling of $\mathbf{x}$ to compute bootstrap samples $\mathbf{x}^*$
- Computation of bootstrap replication of the estimator $\hat{\theta}^*(b)$ for $b = 1, \cdots, B$
- From replications, standard error $\widehat{\mathrm{se}}_B$, the bias $\widehat{\mathrm{Bias}}_B$ and the confidence interval.
- Non-parametric bootstrap estimations (no prior on F).