Simulation 4

MX

2026-01-03

Chap 8 Bootstrap and Jackknife

8.1 The Bootstrap

Bootstrap can refer to nonparametric bootstrap or parametric bootstrap. Parametric bootstrap Monte Carlo methods that involve sampling from a fully specified probability distribution, such as methods of Chapter 7 are sometimes called parametric bootstrap. Nonparametric bootstrap is the subject of this chapter. In nonparametric bootstrap, the distribution is not specified.

The distribution of the finite population represented by the sample can be regarded as a pseudo-population with similar characteristics as the true population. By repeatedly generating random samples from this pseudo-population (resampling), the sampling distribution of a statistic (properties of an estimator such as bias or standard error) can be estimated.

Suppose $x = (x_1, ..., x_n)$ is an observed random sample from a distribution $F(x)$. A (1) sample X^* is selected from x with prob $P(X^* = x_i) = 1/n$.

Resampling generates a random sample $X_1^*, ..., X_n^*$. The rvs X_i^* are independent and identically distributed according to the empirical cdf (ecdf) of X, F_n. That is, $X^* \sim F_n$.

The empirical cdf of the bootstrap replicates, F_n^*, is an approximation to F_n. Resampling from the sample x (bootstrap) is equivalent to random sampling from F_n.

The empirical cdf of X, F_n, approximates F, and the empirical cdf of bootstrap replicates, F_n^*, approximates F_n: F_n^* -> F_n -> F. **Example 8.1**

Suppose θ is the param of interest, and $\hat{\theta}$ is an estimator of θ. Then the bootstrap estimate of the distr of $\hat{\theta}$ $F_{\hat{\theta}}$ is obtained as follows:

1. For each bootstrap replicate, indexed by b = 1,. . . , B: Generate sample $x^{*(b)} = x_1^{*(b)}, ..., x_n^{*(b)}$ by sampling with replacement from the original sample $x = x_1, ..., x_n$. Compute the b^{th} replicate, $\hat{\theta}^{(b)}$, from the bootstrap sample.
2. The bootstrap estimate of $F_{\hat{\theta}}$ is is empirical distr of the replicates $\hat{\theta}^{(1)}, ..., \hat{\theta}^{(b)}, ..., \hat{\theta}^{(B)}$.

8.1.1 Bootstrap Estimation of Standard Error

Let $\overline{\hat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$. The bs estimate of standard error of an estimator $\hat{\theta}$ is the standard deviation of bs replicates $\hat{\theta}^{(1)}, ..., \hat{\theta}^{(b)}, ..., \hat{\theta}^{(B)}$.

$$\hat{se}(\hat{\theta}^*) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{(b)} - \overline{\hat{\theta}^*})}$$

Notation	Description
$x = (x_1, \ldots, x_n)$	Original observed sample
$x^* = (x_1^*, \ldots, x_n^*)$	A generic bootstrap sample
$x^{*(b)}$	The b-th bootstrap sample
$\hat{\theta}$	Estimator
$\hat{\theta}^*$	Generic bootstrap version of the estimator
$\hat{\theta}^{(b)}$	The b-th bootstrap realization of the estimator

Table 1: Notation Summary

B=50 is usually large enough for good estimates of s.e., rarely is B>200 necessary. Much larger B will be needed for confidence interval estimation.

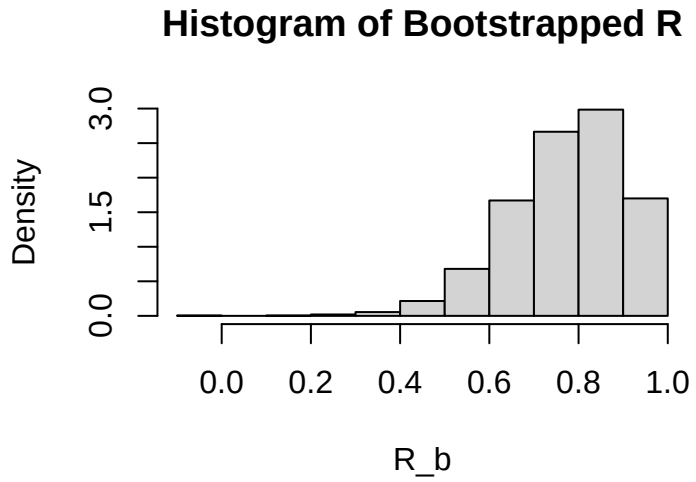
Example 8.2. Estimate the correlation between LSAT and GPA scores, and compute the bs estimate of the standard error of the sample correlation R.

```
library(bootstrap)
data(law) # 15 law schools
R = cor(law$GPA, law$LSAT)  # correlation statistic
cat('correlation of scores:', R)
```

```
## correlation of scores: 0.7763745
```

```
B = 2000
n = nrow(law)  # 15
R_b = numeric(B)
for (b in 1:B){
  index = sample(1:n, size=n, replace=T)
  law_bs = law[index,]   # resample the data
  R_b[b] = cor(law_bs$LSAT, law_bs$GPA) # theta^hat(b)
}

hist(R_b, prob=T, main='Histogram of Bootstrapped R')
```



```r
se_R = sd(R_b)  # bs estimate of the se of R
cat('Bootstarp estimate of se(R):', se_R)
```

```
## Bootstarp estimate of se(R): 0.1285398
```

8.1.2 Bootstrap Estimation of Bias

The bias of an estimator $\hat{\theta}$ for θ is $bias(\hat{\theta}) = E[\hat{\theta} - \theta] = E(\hat{\theta}) - \theta$. (As an example the ML estimator of variance is $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 := \hat{\sigma}^2$, whose bias is $E(\hat{\sigma}^2) - \sigma^2 = E[\frac{n-1}{n} \cdot \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$.)

Let $\overline{\hat{\theta}^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$, and $\hat{\theta} = \hat{\theta}(x)$ (an estimate computed from the original sample). The bootstrap estimate of bias is:

$$\widehat{bias}(\hat{\theta}) = \overline{\hat{\theta}^*} - \hat{\theta}$$

The distr of $\hat{\theta} - \theta$ approximates that of $\hat{\theta}^* - \hat{\theta}$. Therefore the mean of the latter, $E[\hat{\theta}^*] - \hat{\theta}$, should be close to the mean of the former, which is defined as bias. Use $\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$ to estimate $E[\hat{\theta}^*]$, hence the formula.

Why use $\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$ to estimate $E[\hat{\theta}^*]$? For the finite population $x = (x_1, ..., x_n)$ the param is $\hat{\theta}$ (given x) and there are B i.i.d (given x) estimators $\hat{\theta}^{(b)}$. The mean of them is unbiased for $E[\hat{\theta}^*]$.

Ex 8.4. In the law school data of Example 8.2, compute bs estimate of bias in the sample correlation.

```r
bias = mean(R_b) - R
cat('bias:', bias)
```

```
## bias: -0.002556851
```

8.2 The Jackknife

8.3 Bootstrap Confidence Intervals

8.3.1 The Standard Normal Bootstrap CI

If $\hat{\theta}$ is normal or $\hat{\theta}$ is the sample mean and the sample size is large (CLT) then $Z = \frac{\hat{\theta} - E(\hat{\theta})}{se(\hat{\theta})} = \frac{\hat{\theta} - bias - \theta}{se(\hat{\theta})} \sim$ N(0,1) approximately. Here we treat se and the bias as fixed quantities.

One can derive $P(\hat{\theta} - bias - se \cdot z_{\alpha/2} < \theta < \hat{\theta} - bias + se \cdot z_{1-\frac{\alpha}{2}}) = 1 - \alpha$.

If $\hat{\theta}$ is unbiased ($bias = 0$), we get the usual approximate $100(1-\alpha)\%$ CI of θ:

$$(\hat{\theta} - se \cdot z_{\alpha/2}, \hat{\theta} + se \cdot z_{1-\frac{\alpha}{2}})$$

(One would use bootstrap to get se of the estimator; see 8.1.1; this leads to 8.3.4 The Bootstrap t Interval.)

However if $bias \neq 0$, we will need to estimate it. The estimator of bias, another rv, will break the normality. That is, $\frac{\hat{\theta} - \widehat{bias} - \theta}{se(\hat{\theta})}$ is no longer approximately N(0,1).

8.3.2 The Basic Bootstrap CI

The basic bootstrap CI transforms the distribution of the replicates $\hat{\theta}^{(b)}$ by subtracting the observed statistic $\hat{\theta}$. The quantiles of the transformed sample $\hat{\theta^*} - \hat{\theta}$ are used to determine the confidence limits. Specifically, let l and u respectively be the $\alpha/2$ and $1-\alpha/2$ percentile of the distr of $\hat{\theta^*} - \hat{\theta}$:

$$l = \hat{\theta}^*_{\alpha/2} - \hat{\theta}$$

$$u = \hat{\theta}^*_{1-\alpha/2} - \hat{\theta}$$

Then,

$$P(l < \hat{\theta^*} - \hat{\theta} < u) = 1 - \alpha$$

Because the distr of $\hat{\theta} - \theta$ approximates that of $\hat{\theta^*} - \hat{\theta}$,

$$P(l < \hat{\theta} - \theta < u) = 1 - \alpha$$

That is,

$$P(\hat{\theta} - u < \theta < \hat{\theta} - l) = 1 - \alpha$$

Hence the CI

$$(\hat{\theta} - u, \hat{\theta} - l) = (2\hat{\theta} - \hat{\theta}^*_{1-\alpha/2}, 2\hat{\theta} - \hat{\theta}^*_{\alpha/2})$$

8.3.3 The Percentile Bootstrap CI

Uses the quantiles of the empirical distr of the bootstrap replicates directly:

$$(\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2})$$

The quantiles are estimators of the quantiles of the sampling distribution of $\hat{\theta}$, so that these (random) quantiles may match the true distribution better when the distribution of $\hat{\theta}$ is not normal (the standard normal CI).

Example 8.10. Compute 95% bootstrap CI estimates for the correlation statistic in the law data of Example 8.2

```
library(boot)
data(law, package = 'bootstrap')
boot_obj = boot(data=law, R=2000,
                statistic=function(x,i){
                  cor(x[i,1], x[i,2])
                  })
boot.ci(boot.out = boot_obj, type=c('basic', 'norm', 'perc'))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_obj, type = c("basic", "norm", "perc"))
##
## Intervals :
## Level      Normal              Basic              Percentile
## 95%   ( 0.5238,  1.0433 )   ( 0.5938,  1.1015 )   ( 0.4512,  0.9590 )
## Calculations and Intervals on Original Scale
```

```r
# standard normal
l = R - se_R*1.96
u = R + se_R*1.96
cat('std normal CI:', l, u)
```

```
## std normal CI: 0.5244365 1.028312
```

```r
# basic bs CI
l = 2*R - quantile(R_b, probs = 0.975)
u = 2*R - quantile(R_b, probs = 0.025)
cat('\nbasic bs CI:', l, u)
```

```
##
## basic bs CI: 0.5902857 1.069362
```

```r
# percentile bs CI
cat('\npercentile bs CI:', quantile(R_b, probs = 0.025), quantile(R_b, probs = 0.975))
```

```
##
## percentile bs CI: 0.4833866 0.9624633
```

8.3.4 The Bootstrap t Interval

8.4 Better Bootstrap CIs