Computational Statistics. Chapter 6: Bootstrapping. Solution of exercises

Thierry Denoeux

2023-01-28

```
set.seed(20211015)
```

Exercise 1

Question a

Data generation:

```
n<-50
rate<-0.5
x<-rexp(n,rate=rate)
```

The cdf of the exponential distribution is $F(x) = 1 - \exp(-\theta x)$. The equation $F(m) = 0.5$ gives us $m = \log(2)/\theta$ Here, we get

```
mtrue<-log(2)/rate
print(mtrue)
```

```
## [1] 1.386294
```

Question b

```
bootstrap <- function(x,B){
    n<-length(x)
    X<-matrix(0,B,n)
    for(b in (1:B)) X[b,]<-sample(x,size=n,replace=TRUE)
    return(X)
}
```

Question c

The MLE of θ is $\widehat{\theta} = 1/\overline{X}$, where $\overline{X}$ is the sample mean. The plug-in estimate of m is

$$\widehat{m} = \log(2)/\widehat{\theta} = \log(2)\overline{X}.$$

We write the following R function:

```r
median_plug<-function(x) return(mean(x)*log(2))
```

For our dataset, we get

```r
mhat<-median_plug(x)
print(mhat)
```

```
## [1] 1.552819
```

To estimate the standard error of $\widehat{m}$, we generate B bootstrap samples and we compute the standard deviation of the bootstrap estimates:

```r
B<- 1000
X<-bootstrap(x,B)  # generation of B bootstrap samples
med<-apply(X,1,median_plug)  # calculation of the B bootstrap estimates
se_B<- sd(med)  # calculation of the standard deviation
print(se_B)
```

```
## [1] 0.1806734
```

Question d

The bounds of the $1-\alpha$ bootstrap percentile confidence interval are simply the $\alpha/2$ anf $1-\alpha/2$ empirical quantiles of the B bootstrap estimates:

```r
alpha<- 0.05
CI1<-quantile(med,c(alpha/2,1-alpha/2))
print(CI1)
```

```
##     2.5%    97.5%
## 1.221340 1.912688
```

Question e

To compute bootstrap estimates of the standard error of $\widehat{m}$, we need to bootstrap each bootstrap sample. We will use only $B_1 = 50$ pseudo-datasets in the inner bootstrap loop:

```r
B1<-50
se<-vector("numeric",B)
for(b in 1:B){
    X1<- bootstrap(X[b,],B1)  # Bootstrapping bootstrap sample b
    med_star<-apply(X1,1,median_plug)
    se[b]<- sd(med_star)
    }
```

We can then compute the quantiles of the approximately pivotal quantity

$$Z^* = \frac{\widehat{m}^* - \widehat{m}}{\mathsf{se}^*}$$

```r
q<-quantile((med-mhat)/se,c(alpha/2,1-alpha/2))
CI2<- c(mhat-q[2]*se_B,mhat-q[1]*se_B)
print(CI2)
```

```
##    97.5%     2.5%
## 1.213185 2.022918
```

Question f

To estimate the coverage probabilities of the two confidence intervals above, we will generate $N = 200$ datasets and compute the corresponding confidence intervals. The coverage probability is estimated by the proportion of intervals containing the true value of the parameter.

```
N<-200
k1<-0
k2<-0
for(i in 1:N){
  x<-rexp(n,rate=rate)  # Sample generation
  X<-bootstrap(x,B)  # bootstrapping
  med<-apply(X,1,median_plug)  # computation of the plug-in estimates
  CI1<-quantile(med,c(alpha/2,1-alpha/2))  # percentile confidence interval
  if((CI1[1]<=mtrue)&(mtrue<=CI1[2])) k1<-k1+1  # k1 is incremented if the CI contains the true median
  mhat<-median_plug(x)
  for(b in 1:B){
    X1<- bootstrap(X[b,],B1)  # Bootstrapping bootstrap sample b
    med_star<-apply(X1,1,median_plug)
    se[b]<- sd(med_star)
  }
  q<-quantile((med-mhat)/se,c(alpha/2,1-alpha/2))
  CI2<- c(mhat-q[2]*se_B,mhat-q[1]*se_B)  # Bootstrap t interval
  if((CI2[1]<=mtrue)&(mtrue<=CI2[2])) k2<-k2+1
}
p1<-k1/N
p2<-k2/N
print(c(p1,p2))
```

```
## [1] 0.955 0.960
```

The two confidence intervals appear to have similar coverage probabilities.

Exercise 2

Question a

We first read and transform the data, and create a new data frame:

```
data <- read.table("/Users/Thierry/Documents/R/Data/Compstat/investment.txt", header=TRUE)
y<- data$Invest/(data$CPI*10)
time<- 1:15
GNP1<-data$GNP/(data$CPI*10)
data1<-data.frame(Invest=y,time, GNP=GNP1,data$Interest,data$Inflation)
```

We then perform linear regression on this dataset using function `lm`, and compute 95% confidence intervals on the coefficients under the assumption of the standard linear regression model:

```
reg<- lm(Invest~.,data=data1)
CI.norm<-confint(reg, level = 0.95)
print(CI.norm)
```

```
##                       2.5 %        97.5 %
## (Intercept)    -0.629244578 -0.3888960778
## time           -0.020888726 -0.0122906337
```