Advanced Statistical Methods Hw8

Do Hyup Shin

2021-11-24

Problem 10.4

Verify formula (10.38) for the number of distinct bootstrap samples.

Solution

We'll show that the number of distinct bootstrap samples = $\binom{2n-1}{n}$. This problem is a duplicate combination problem. Let $(x_1, x_2, \ldots, x_n)$ be the sample and the size of sample is n. Let the number of times each observation is chosen is a_i $\forall i = 1, 2, \ldots, n$. Then, $\sum_{i=1}^{n} a_i = n$ with $\forall 0 \leq a_i \leq n$ and $\forall a_i$ are nonnegative integer. We should find the number of combination a_i satisfying above condition. This problem is the same as following problem. Suppose that there exist n-1 bars(= $|$) and n dots(= $\cdot$). Let's arrange the two types of symbols in a row. Then, we can express the arranged line in this way ___ | ___ | ___ . . . ___ | ___ | ___ and ___ means where $\cdot$ can enter. There exists n seperation which is ___.

Thus, we can correspond $\forall a_i$ to the number of $\cdot$ in ith ___. We know that the number of permutations n-1 bars(= $|$) and n dots(= $\cdot$) is $\dfrac{(2n-1)!}{(n-1)!n!} = \binom{2n-1}{n}$.

Therefore, the number of distinct bootstrap samples is $\binom{2n-1}{n}$.

Problem 10.5

A normal theory least squares model (7.28)-(7.30) yields $\hat{\beta}$ (7.32). Describe the parametric bootstrap estimates for the standard errors of the components of $\hat{\beta}$.

Solution

$Y = X\beta + \epsilon$ and $Y - X\beta = \epsilon \sim N_n(0, I_n\sigma^2)$. The mle of β is $\hat{\beta} = (X^TX)^{-1}X^TY$ and $\hat{\sigma}^2 = MSE = \frac{1}{n-p-1}Y^T(I-H)Y$. Then, we plug in $\hat{\beta}$ and $\hat{\sigma}^2$ instead of β and σ^2. Then we can use bootstrap sampling $y_i - x_i^t\hat{\beta} = \epsilon_i^* \overset{iid}{\sim} N(0, \hat{\sigma}^2)$ $\forall i = 1, 2, \ldots, n$. Then, we define the new regression model such that $y_i^* = x_i^t\hat{\beta} + \epsilon_i^*$. Let $Y^* = (y_1^*, y_2^*, \ldots, y_n^*)^T$. In this model, we can regress X on Y^*, so $\hat{\beta}^* = (X^TX)^{-1}X^TY^*$.

By above process, Some large number B of bootstrap samples are independently drawn. The corresponding bootstrap replications are calculated, say $\hat{\beta}^{*b} = (X^TX)^{-1}X^TY^{b*}$. Therefore, we can estimate the bootstrap standard error of β_i such that

$$\hat{se}_{boot}(\hat{\beta}_i) = (\frac{1}{B-1}\sum_{j=1}^{B}({\hat{\beta}_i}^{*j} - \hat{\beta}_{i(\cdot)})^2)^{1/2}$$

where $\hat{\beta}_{i(\cdot)} = \frac{1}{B}\sum_{j=1}^{B}\hat{\beta}_i^{*j}$

Problem 10.7

Verify formula (10.70).

Solution

We'll show that the variance of sample mean of bootstrap sample $X^* = (x_1^*, x_2^*, \dots, x_n^*)$ is $\sum_{i=1}^n (x_i - \bar{x})^2/n^2$. Let $X = (x_1, x_2, \dots, x_n)$ be random sample from popuplation F and define $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Let $\hat{F}$ be the empirical probability distribution that puts probability 1/n on each point x_i. The bootstrap sample with replace from $\{x_1, x_2, \dots, x_n\}$ is $X^* = (x_1^*, x_2^*, \dots, x_n^*)$, i.e $x_i^* \overset{iid}{\sim} \hat{F}$. Then, $P(x_i^* = x_j) = \frac{1}{n} \quad \forall 1 \le i, j \le n$. So the expectation of x_i^* is $E_{\hat{F}}(x_i^*) = \sum_{j=1}^n x_j P(x_i^* = x_j) = \sum_{j=1}^n x_j \frac{1}{n} = \frac{1}{n}\sum_{j=1}^n x_j = \bar{x}$.

Define $\bar{x}^* = \frac{1}{n}\sum_{j=1}^n x_j^*$ which is sample mean of bootstrap sample. Then, the variance of $\bar{x}^*$ is

$$
\begin{aligned}
var_{\hat{F}}(\bar{x}^*) &= E_{\hat{F}}((\bar{x}^* - E(\bar{x}^*))^2) = E_{\hat{F}}((\bar{x}^* - \bar{x})^2) \\
&= E_{\hat{F}}(\sum_{j=1}^n \frac{1}{n}(x_j^* - \bar{x}))^2 = \frac{1}{n^2} E_{\hat{F}}(\sum_{j=1}^n (x_j^* - \bar{x}))^2 \\
&= \frac{1}{n^2} E_{\hat{F}}(\sum_{j=1}^n (x_j^* - \bar{x})^2 + \sum_{i \ne j}(x_i^* - \bar{x})(x_j^* - \bar{x})) \\
&= \frac{1}{n^2} E_{\hat{F}}(\sum_{j=1}^n (x_j^* - \bar{x})^2) \quad (\because E_{\hat{F}}(x_i^* - \bar{x})(x_j^* - \bar{x}) = 0 \; \forall i \ne j) \\
&= \frac{1}{n^2} n E_{\hat{F}}(x_1^* - \bar{x})^2 \quad (\because \forall (x_j^* - \bar{x}) \text{ are following independently identical distribution}) \\
&= \frac{1}{n}\sum_{j=1}^n (x_j - \bar{x})^2 P(x_1^* = x_j) = \frac{1}{n}\sum_{j=1}^n \frac{1}{n}(x_j - \bar{x})^2 \\
&= \frac{1}{n^2}\sum_{j=1}^n (x_j - \bar{x})^2
\end{aligned}
$$

Therefore, $var_{\hat{F}}(\bar{x}^*) = \frac{1}{n^2}\sum_{j=1}^n (x_j - \bar{x})^2$. Suppose that there exist B bootstrap samples. Then, We can calculate $\bar{x}^{*j}$ for each jth bootstrap sample. So the estimate of $var_{\hat{F}}(\bar{x}^*)$ is $v\hat{a}r_{boot}(\bar{x}^*) = \frac{1}{B-1}\sum_{j=1}^B (\bar{x}^{*j} - \bar{x}_{(\cdot)})^2$ where $\bar{x}_{(\cdot)} = \frac{1}{B}\sum_{j=1}^B \bar{x}^{*j}$.

In conclusion, $v\hat{a}r_{boot}(\bar{x}^*) = \frac{1}{B-1}\sum_{j=1}^B (\bar{x}^{*j} - \bar{x}_{(\cdot)})^2 \to var_{\hat{F}}(\bar{x}^*) = \frac{1}{n^2}\sum_{j=1}^n (x_j - \bar{x})^2$ as $B \to \infty$.

Problem 10.9

A survey in a small town showed incomes $x_1, x_2, \dots, x_m$ for men and $y_1, y_2, \dots, y_n$ for women. As an estimate of the differences,

$$\hat{\theta} = median\{x_1, x_2, \dots, x_m\} - median\{y_1, y_2, \dots, y_n\}$$

was computed.

(a) How would you use nonparametric bootstrapping to assess the accuracy of $\hat{\theta}$?
(b) Do you think your method makes full use of the bootstrap replications?

Solution

(a)

Let $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_n)$ be the samples of men and women, respectively. Some large number B of bootstrap samples are independently drawn. Let $X^{*j} = (x_1^{*j}, x_2^{*j}, \dots, x_m^{*j})$ and $Y^{*j} = (y_1^{*j}, y_2^{*j}, \dots, y_n^{*j})$ be the Bth bootstrap sample of X and Y, respectively. The corresponding bootstrap replications are calculated, say $\hat{\theta}^{*j} = median\{x_1^{*j}, x_2^{*j}, \dots, x_m^{*j}\} - median\{y_1^{*j}, y_2^{*j}, \dots, y_n^{*j}\}$. Then, the bootstrap estimate of standard error for $\hat{\theta}$ is $\hat{se}_{boot}(\hat{\theta}) = (\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}^{*i} - \hat{\theta}_{(\cdot)})^2)^{1/2}$. So, we can assess the accuracy of $\hat{\theta}$ by above process.

(b)

Problem 11.1

We observe $y \sim \lambda G_{10}$ to be $y = 20$. Here λ is an unknown parameter while G_{10} represents a gamma random variable with 10 degrees of freedom ($y \sim G(10, \lambda)$ in the notation of Table 5.1). Apply the Neyman constructions as in Figure 11.1 to find the confidence limit endpoints $\hat{\lambda}(0.025)$ and $\hat{\lambda}(0.975)$.

Solution

The pdf of y is $f_\lambda(y) = \frac{1}{\Gamma(10)\lambda^{10}} y^9 e^{-\frac{y}{\lambda}}$. The loglikelihood function is $l(\lambda) = log(f_\lambda(y)) = -log9! - 10log\lambda + 9logy - \frac{y}{\lambda}$. We can find the mle of λ satisfying $\frac{\partial l}{\partial \lambda} = -\frac{10}{\lambda} + \frac{y}{\lambda^2} = 0$. Then, the mle of λ is $\hat{\lambda} = \frac{y}{10} = 2$. Since $y \sim G(10, \lambda)$, $\hat{\lambda} = \frac{y}{10} \sim G(10, \frac{\lambda}{10})$.

We'll show that $\lambda_1 \le \lambda_2 \Rightarrow P_{\lambda_2}(\hat{\lambda} \le r) \le P_{\lambda_1}(\hat{\lambda} \le r)$. The cdf of $\hat{\lambda}$ is

$$\begin{aligned}
F_\lambda(r) &= \int_0^r \frac{1}{9!(\lambda/10)^{10}} x^9 e^{-\frac{10x}{\lambda}} dx \\
&= \int_0^{10r/\lambda} \frac{1}{9!} t^9 e^{-t} dx \quad (10x/\lambda = t, dx = \lambda/10dt) \\
&= \frac{1}{9!} \int_0^{10r/\lambda} t^9 e^{-t} dx
\end{aligned}$$

Then, $F_\lambda(r)$ is decreasing function of λ because the integral interval $(0, 10r/\lambda)$ is reduced when λ is incresing. Thus, $\lambda_1 \le \lambda_2 \Rightarrow F_{\lambda_2}(r) \le F_{\lambda_1}(r)$. Define the function of α-quantile of $\hat{\lambda}$ for λ denoted $g_\alpha(f_\lambda)$ satisfying $P_\lambda(\hat{\lambda} \le g_{\frac{\alpha}{2}}(f_\lambda)) = \frac{\alpha}{2}$. Then, $g_\alpha(f_\lambda)$ is increaing function for λ.

So, we'll find $\hat{\lambda}_{(up)}$ and $\hat{\lambda}_{(lo)}$ such that $g_{0.025}(f_{\hat{\lambda}_{(up)}}) = \hat{\lambda}$ and $g_{0.975}(f_{\hat{\lambda}_{(lo)}}) = \hat{\lambda}$. This means $P_{\hat{\lambda}_{lo}}(X \ge \hat{\lambda}) = \int_{\hat{\lambda}}^{\infty} f_{\hat{\lambda}_{lo}}(x)dx = 0.025$ and $P_{\hat{\lambda}_{up}}(X \le \hat{\lambda}) = \int_0^{\hat{\lambda}} f_{\hat{\lambda}_{up}}(x)dx = 0.025$ where $f_\lambda(x)$ is the pdf of $\hat{\lambda}$.

We know that $\hat{\lambda} \sim G(10, \frac{\lambda}{10}) \Leftrightarrow \frac{20}{\lambda}\hat{\lambda} \sim G(10, 2) = \chi^2(20)$. Using this, $\frac{20}{\hat{\lambda}_{(up)}}\hat{\lambda} \sim \chi^2(20)$ and $\frac{20}{\hat{\lambda}_{(lo)}}\hat{\lambda} \sim \chi^2(20)$

respectively. Then,

$$
\begin{aligned}
P_{\hat{\lambda}_{lo}}(X \geq \hat{\lambda}) &= P_{\hat{\lambda}_{lo}}(\frac{20}{\hat{\lambda}_{(lo)}}X \geq \frac{20}{\hat{\lambda}_{(lo)}}\hat{\lambda}) \\
&= P(Y \geq \frac{20}{\hat{\lambda}_{(lo)}}\hat{\lambda}) \quad (Y = \frac{20}{\hat{\lambda}_{(lo)}}X \sim \chi^2(20)) \\
&= 0.025
\end{aligned}
$$

So, $\frac{20}{\hat{\lambda}_{(lo)}}\hat{\lambda} = \chi^2_{0.025}(20) \to \hat{\lambda}_{(lo)} = \frac{20\hat{\lambda}}{\chi^2_{0.025}(20)} = \frac{40}{\chi^2_{0.025}(20)}$

```r
#mle of lambda
hat_lambda = 2
#the value of lambda_lo
hat_lam_lo = 20*hat_lambda/qchisq(0.025, 20, lower.tail = F)
hat_lam_lo
```

```
## [1] 1.170631
```

$\hat{\lambda}_{(lo)} = \frac{40}{\chi^2_{0.025}(20)} = 1.170631.$

Similarly,

$$
\begin{aligned}
P_{\hat{\lambda}_{up}}(X \leq \hat{\lambda}) &= P_{\hat{\lambda}_{up}}(\frac{20}{\hat{\lambda}_{(up)}}X \leq \frac{20}{\hat{\lambda}_{(up)}}\hat{\lambda}) \\
&= P(Y \leq \frac{20}{\hat{\lambda}_{(up)}}\hat{\lambda}) \quad (Y = \frac{20}{\hat{\lambda}_{(up)}}X \sim \chi^2(20)) \\
&= 0.025
\end{aligned}
$$

Thus, $\frac{20}{\hat{\lambda}_{(up)}}\hat{\lambda} = \chi^2_{0.975}(20) \to \hat{\lambda}_{(up)} = \frac{20\hat{\lambda}}{\chi^2_{0.975}(20)} = \frac{40}{\chi^2_{0.975}(20)}$.

```r
#the value of lambda_up
hat_lam_up = 20*hat_lambda/qchisq(0.975, 20, lower.tail = F)
hat_lam_up
```

```
## [1] 4.170673
```

$\hat{\lambda}_{(up)} = \frac{40}{\chi^2_{0.975}(20)} = 4.170673.$

Therefore, $\hat{\lambda}(0.025) = 1.170631$ and $\hat{\lambda}(0.975) = 4.170673$.

The 95% confidence interval of Neyman constructions is (1.170631, 4.170673).

Problem 11.3

Suppose $\hat{G}$ in (11.33) was perfectly normal, say $\hat{G} \sim N(\hat{\mu}, \hat{\sigma}^2)$. What does $\hat{\theta}_{BC}(\alpha)$ reduce to in this case, and why does this make intuitive sense?

Solution

Suppose that $\hat{G}$ is cdf of $N(\hat{\mu}, \hat{\sigma}^2)$ with $z_0 = \Phi^{-1}(p_0)$ and $z^{(\alpha)} = \Phi^{-1}(\alpha)$ where Φ is cdf of standard normal distribution. Also, $p_0 = \frac{\#\{\hat{\mu}^{*b} \leq \hat{\mu}\}}{B}$ and $z_0 = \Phi^{-1}(p_0)$. Therefore,

$$\hat{\theta}_{BC}[\alpha] = \hat{G}^{-1}[\Phi(2z_0 + z^{(\alpha)})]$$

Since $\hat{G}(t) = \Phi(\dfrac{t - \hat{\mu}}{\hat{\sigma}})$, $\hat{G}(\hat{\theta}_{BC}[\alpha]) = \Phi(\frac{\hat{\theta}_{BC}[\alpha] - \hat{\mu}}{\hat{\sigma}}) = \Phi(2z_0 + z^{(\alpha)})$.

By solving above equation, $\hat{\theta}_{BC}[\alpha] = \hat{\mu} + \hat{\sigma}(2z_0 + z^{(\alpha)})$. If $B \to \infty$, then $p_0 \approx 0.5$ and $z_0 = \Phi^{-1}(p_0) \approx 0$. So, $\hat{\theta}_{BC}[\alpha] = \hat{\mu} + \hat{\sigma}(2z_0 + z^{(\alpha)}) \approx \hat{\mu} + \hat{\sigma} z^{(\alpha)}$.

Thus, if $\hat{G}$ is normal, bias-corrected confidence interval is almost the same as standard interval.

Problem 11.5

Suppose $\hat{\theta} \sim Poisson(\theta)$ is obeserved to equal 16. Without employing simulation, compute the 95% central BCa interval for θ. (You can use the good approximation $z_0 = a = 1/(6\hat{\theta}^{1/2})$.)

Solution

Let $\hat{G}(t) = \frac{\#\{\hat{\theta}^{*b} \leq t\}}{B}$ and $p_0 = \hat{G}(\hat{\theta})$ and $z_0 = \Phi^{-1}(p_0) = a = 1/(6\hat{\theta}^{1/2}) = \frac{1}{24}$. We'll find $\hat{\theta}_{BCa}[0.025]$ and $\hat{\theta}_{BCa}[0.975]$. We know that $z^{(0.025)} = -1.96$ and $z^{(0.975)} = 1.96$. First, the value of $\hat{\theta}_{BCa}[0.025]$ is

$$\hat{\theta}_{BCa}[0.025] = \hat{G}^{-1}[\Phi(z_0 + \frac{z_0 + z^{(0.025)}}{1 - a(z_0 + z^{(0.025)})})] = \hat{G}^{-1}[\Phi(\frac{1}{24} + \frac{1/24 - 1.96}{1 - 1/24(1/24 - 1.96)})] = \hat{G}^{-1}[\Phi(-1.735)] = \hat{G}^{-1}(0.041)$$

Second, the value of $\hat{\theta}_{BCa}[0.975]$ is

$$\hat{\theta}_{BCa}[0.975] = \hat{G}^{-1}[\Phi(z_0 + \frac{z_0 + z^{(0.975)}}{1 - a(z_0 + z^{(0.975)})})] = \hat{G}^{-1}[\Phi(\frac{1}{24} + \frac{1/24 + 1.96}{1 - 1/24(1/24 + 1.96)})] = \hat{G}^{-1}[\Phi(2.225)] = \hat{G}^{-1}(0.987)$$

Therefore, the 95% BCa interval for θ is $(\hat{G}^{-1}(0.041),\ \hat{G}^{-1}(0.987))$.

If $B \to \infty$, $\hat{G}(t) \xrightarrow{p} P(X \leq t)$ where $X \sim Poisson(16)$. Thus, we can find the $\hat{G}^{-1}(0.041)$ and $\hat{G}^{-1}(0.987)$ by using qpois.

```
qpois(0.041, 16)
```

```
## [1] 9
```

```
qpois(0.987, 16)
```

```
## [1] 26
```

Then, $\hat{G}^{-1}(0.041) = 9$ and $\hat{G}^{-1}(0.987) = 26$, the 95% BCa interval for θ is (9, 26).

Problem 11.6

Use the R program bcajack (available with its help file from efron.web.stanford.edu under "Talks") to find BCa confidence limits for the student score eigenratio statistic as in Figure 10.2.

Solution

By using "bcajack" function in "bcaboot" package, we can find the BCa confidence interval for the student score eigenratio.

```
library(bcaboot)

#Read the student score data
stu_score <- read.csv("https://web.stanford.edu/~hastie/CASI_files/DATA/student_score.txt", sep = " ")

#original eigen ratio
cor_mat_sco <- cor(stu_score)
eigen_ratio = max(eigen(cor_mat_sco)$values) / sum(eigen(cor_mat_sco)$values)

#eigen_ratio function
eigen_ratio_func <- function(x){
  cor_x = cor(x)
  eigen_ratio_x = max(eigen(cor_x)$values) / sum(eigen(cor_x)$values)
  return(eigen_ratio_x)
}

set.seed(1234)
bca_inter = bcajack(x = stu_score, B = 2000, func = eigen_ratio_func , m = 10, verbose = FALSE)
```

```
## Warning in 2 * t. - s.: longer object length is not a multiple of shorter object
## length
```

```
bca_inter
```

```
## $call
## bcajack(x = stu_score, B = 2000, func = eigen_ratio_func, m = 10,
##     verbose = FALSE)
##
## $lims
##             bca      jacksd       std    pct
## 0.025 0.5274918 0.009346084 0.5441957 0.0315
## 0.05  0.5556361 0.011652410 0.5680448 0.0550
## 0.1   0.5893499 0.006613403 0.5955413 0.0995
## 0.16  0.6116176 0.003264221 0.6172699 0.1515
## 0.5   0.6892780 0.002321084 0.6925353 0.4640
## 0.84  0.7587103 0.004298275 0.7678007 0.8300
## 0.9   0.7796605 0.003585370 0.7895294 0.8995
## 0.95  0.8018533 0.003296287 0.8170258 0.9555
## 0.975 0.8203374 0.004612939 0.8408749 0.9815
##
## $stats
##         theta      sdboot          z0          a     sdjack
## est 0.6925353 0.075684847 -0.04513463 0.05388811 0.03619517
## jsd 0.0000000 0.001475868  0.03039936 0.00000000 0.00000000
##
## $B.mean
## [1] 2000.0000000    0.6879831
##
## $ustats
##      ustat        sdu
## 0.69708749 0.08359081
```

```
##
## attr(,"class")
## [1] "bcaboot"
```

Therefore, the 95% BCa confidence interval for eigenratio is (0.5275, 0.8203).