Stat 3005 – Nonparametric Statistics

Assignment 6

Due: 2 December (Mon) at 1800

Fall 2024

Exercise 6.1 (★★☆ — Modified bootstrap procedures for multivariate data (100%))**.** Revisit the cloud seeding problem in Assignment 5. In this exercise, we further assume that $X_i(a) = \mu_a + \varepsilon_{ia}$ for $a = 0, 1$ and $i = 1, \ldots, n$, where ε_{ia}'s are identically distributed with mean zero. Consider two estimators for θ:

$$\widehat{\theta} = \frac{2}{n}\sum_{i=1}^{n}\{X_iA_i - X_i(1-A_i)\} \qquad \text{and} \qquad \widetilde{\theta} = \frac{\sum_{i=1}^{n} X_iA_i}{\max\{1, \sum_{i=1}^{n} A_i\}} - \frac{\sum_{i=1}^{n} X_i(1-A_i)}{\max\{1, \sum_{i=1}^{n}(1-A_i)\}}.$$

1. (10%) The dataset $Y_i = (A_i, X_i)$, $i = 1, \ldots, n$, is bivariate. So, the univariate R functions (`var.jack`, `var.boot`, `ci.pboot`, `ci.spboot`, and `ci.bcaboot`) in the lecture note are not directly implementable. But they can be modified very easily. For example, `var.jack` can be modified as follows.

```
var.jack = function(x, FUN){
    x = as.matrix(x)              # Modification 1: redefine x as a matrix
    n = nrow(x)                   # Modification 2: change length(x) to nrow(x)
    out = rep(NA,n)
    for(i in 1:n){
        out[i] = FUN(x[-i,])      # Modification 3: change x[-i] to x[-i,]
    }
    (n-1)*(mean(out^2)-mean(out)^2)
}
```

   Rewrite the functions `var.boot`, `ci.pboot`, `ci.spboot`, and `ci.bcaboot` so that they can handle general multivariate data input.

2. Variance estimation
   (a) (10%) Briefly explain why $\widetilde{\theta}$ is a sensible estimator when n is large enough. (Use $\lesssim 50$ words.)
   (b) (10%) Briefly describe the difficulties of deriving the theoretical value of $\mathsf{Var}(\widetilde{\theta})$. (Use $\lesssim 50$ words.)
   (c) (10%) Use jackknife to estimate $\mathsf{Var}(\widehat{\theta})$ and $\mathsf{Var}(\widetilde{\theta})$.
   (d) (10%) Use bootstrap to estimate $\mathsf{Var}(\widehat{\theta})$ and $\mathsf{Var}(\widetilde{\theta})$.
   (e) (10%) Which estimator ($\widehat{\theta}$ or $\widetilde{\theta}$) are better in terms of precision?

3. Construction of confidence intervals
   (a) (10%) Use percentile bootstrap to find a 95% two-sided CI for θ based on $\widetilde{\theta}$.
   (b) (10%) Use studentized pivotal bootstrap to find a 95% two-sided CI for θ based on $\widetilde{\theta}$.
   (c) (10%) Use bias-corrected and accelerated bootstrap to find a 95% two-sided CI for θ based on $\widetilde{\theta}$.
   (d) (10%) Which CI ($\widehat{C}_{\text{P}}$, $\widehat{C}_{\text{SP}}$ or $\widehat{C}_{\text{BCA}}$) do you prefer? Why? (Use $\lesssim 50$ words.)

Hints: See Remark 6.1. Don't read the hints unless you have no ideas and have tried for more than 15 mins.

Exercise 6.2 (★★★ — ✂ Pivotal bootstrap CI)**.** Suppose $T = T(X_1, \ldots, X_n)$ is an estimator of θ based on the sample $X_1, \ldots, X_n$. Let $T^{*(1)}, \ldots, T^{*(B)}$ be B bootstrap replicates of T, as defined in Definition 7.3. Denote the order statistics of $T^{*(1)}, \ldots, T^{*(B)}$ as $T^*_{(1)} \leq T^*_{(2)} \leq \cdots \leq T^*_{(B)}$. The $1-\alpha$ bootstrap pivotal confidence interval is defined as

$$\widehat{C}_{\text{PI}} = \left[2T - T^*_{((1-\alpha/2)B)},\ 2T - T^*_{((\alpha/2)B)}\right].$$

Argue that $\widehat{C}_{\text{PI}}$ is a $1-\alpha$ confidence interval of θ. Only soft argument is needed as in Examples 7.16–7.17.

Remark 6.1 (Hints for Exercise 6.1)**.**
1. Straightforward.
2. (a) What are the limits of $\sum_{i=1}^{n} X_iA_i/n$ and $\max\{1, \sum_{i=1}^{n} A_i\}/n$?
   (b) Finding $\mathsf{Var}(A/B)$ and $\mathsf{Var}(\max(1, B))$ are difficult because ...
   (e) The smaller the variance, the higher the precision.
3. (d) Open-ended.

Instructions: Please follow the instructions stated on the last page of assignment 1.