Computing in Epidemiology and Biostatistics

Modern statistical computing in R: Monte-Carlo simulations (Point estimates)

Wan-Yu Lin

The bias of an estimate: $Bias\left(\hat{\theta}\right)=E\left(\hat{\theta}\right)-\theta$, where θ is the true value in the population, say, population mean. In simulations, we usually use $\frac{\sum_{i=1}^{no.rep}\hat{\theta}_i}{no.rep}$ to estimate $E\left(\hat{\theta}\right)$, where $\hat{\theta}_i$ is the point estimate of the ith replication.

For example, $Bias\left(\bar{X}\right)=E\left(\bar{X}\right)-\mu$

1st replication, generate sample, $\bar{X}_1$

2nd replication, generate sample, $\bar{X}_2$

…

1000th replication, generate sample, $\bar{X}_{1000}$

Use $\frac{\sum_{i=1}^{1000}\bar{X}_i}{1000}$ to estimate $E\left(\bar{X}\right)$.

Ex: Let data come from $N\left(\mu=75,\sigma=15\right)$. If the sample size is 10, please use simulations to evaluate the bias of $\hat{\sigma}^2_{MLE}$ and S^2 (sample variance). Which would be the unbiased estimate of σ^2? Number of replications = 10000, and seed numbers from 1 to 10000, respectively.

Ex: If sleep duration (*Y*) is normally distributed, and

$$Y_i=7-0.05\cdot Age_i+0.5\cdot Watchpc_i+\varepsilon_i,$$

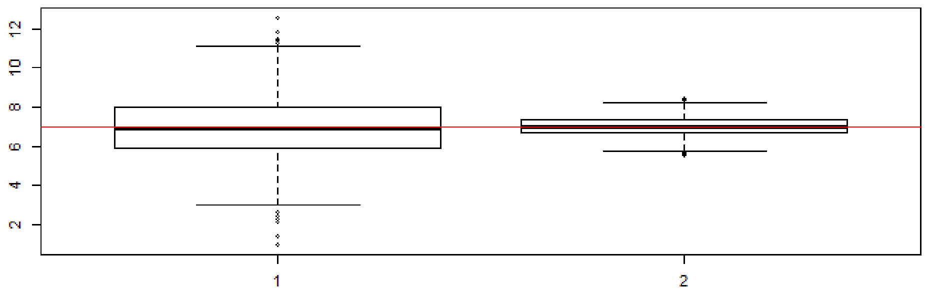
where Age_i is the age of the ith subject, $Watchpc_i$ is the time for watching television of the ith subject, ε_i is the random error term following the standard normal distribution.

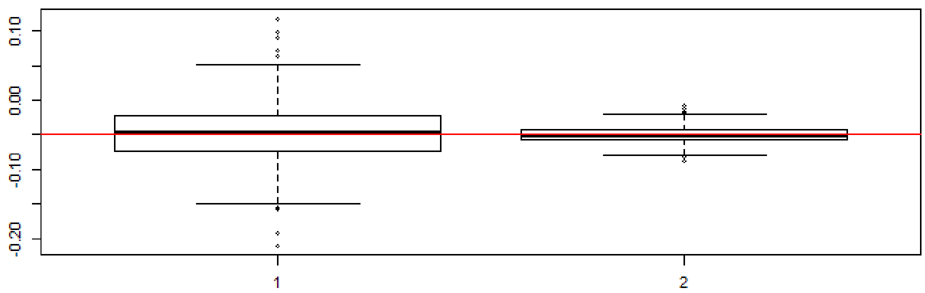
$Age\sim N\left(\mu=35,\sigma=5\right)$, and $Watchpc\sim N\left(\mu=4,\sigma=1\right)$. Suppose the sample sizes are 30 and 300, respectively. Number of replications = 1000, and seed numbers from 1 to 1000, respectively.

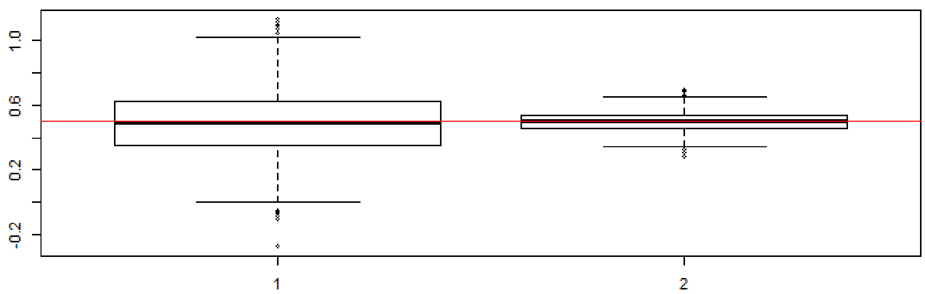
Please use simulations to evaluate the bias of maximum likelihood estimates (MLE) of the regression coefficients, and make box plots to see their distributions.

```
#   (1)   the sample size is 30

betaco <- c(7,-0.05,0.5)
n <- 30
no.rep <- 1000
MLE <- matrix(NA,no.rep,3)        # MLE calculated by ourselves
MLElm <- matrix(NA,no.rep,3)    # MLE calculated by the R built-in function "lm"
for(i in 1:no.rep){
   set.seed(i)
   age <- rnorm(n,35,5)
   watchpc <- rnorm(n,4,1)
   random.error <- rnorm(n,0,1)
   Y <- betaco[1]+betaco[2]*age+betaco[3]*watchpc+random.error
# above: data generation process
# below: data analysis process
   X <- cbind(rep(1,length(Y)),age,watchpc)
   MLE[i,] <- solve(t(X)%*%X)%*%t(X)%*%Y
   MLElm[i,] <- lm(Y~age+watchpc)$coef
}
meanMLE <- colSums(MLE)/no.rep
meanMLElm <- colSums(MLElm)/no.rep

meanMLE
meanMLElm

MLE30 <- MLE

#   (2)   the sample size is 300

n <- 300
no.rep <- 1000
MLE <- matrix(NA,no.rep,3)
MLElm <- matrix(NA,no.rep,3)
for(i in 1:no.rep){
   set.seed(i)
```

```
   age <- rnorm(n,35,5)
   watchpc <- rnorm(n,4,1)
   random.error <- rnorm(n,0,1)
   Y <- betaco[1]+betaco[2]*age+betaco[3]*watchpc+random.error
# above: data generation process
# below: data analysis process
   X <- cbind(rep(1,length(Y)),age,watchpc)
   MLE[i,] <- solve(t(X)%*%X)%*%t(X)%*%Y
   MLElm[i,] <- lm(Y~age+watchpc)$coef
}
meanMLE <- colSums(MLE)/no.rep
meanMLElm <- colSums(MLElm)/no.rep

meanMLE
meanMLElm
MLE300 <- MLE

#    make box plots
par(mfrow = c(3,1))
boxplot(MLE30[,1],MLE300[,1])
abline(h=betaco[1],col=2)
boxplot(MLE30[,2],MLE300[,2])
abline(h=betaco[2],col=2)
boxplot(MLE30[,3],MLE300[,3])
abline(h=betaco[3],col=2)
```







Ex 22: In the above example, please find the coverage of 95% confidence intervals of the regression coefficients when the sample size is 10 and 3000, respectively.

Homework (8 points, please pay attention to all the words in this orange box)

Ex 22-1: If the probability of getting an admission is π_i for the ith subject, which might be related to his/her gpa and gre scores. Considering the model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -6 + gpa_i + 0.005 gre_i$$

$gpa \sim N(\mu = 3.1, \sigma = 0.3)$ and $gre \sim N(\mu = 580, \sigma = 80)$. Suppose the sample sizes are 30, 230, 430, and 630, respectively. Number of replications = 1000, and seed numbers from 1 to 1000, respectively. Please use simulations to evaluate the bias of maximum likelihood estimates (MLE) of the regression coefficients, and make box plots to see their distributions.

Hint:

1. Generate gpa and gre scores for n subjects.
2. Given gpa and gre, claculate π_i.
3. sample(c(0,1),1,c(1-π_i, π_i),replace=F)
4. Use the Newton-Raphson method to find MLEs.

Please note: Using the R built-in function "glm(admit~gpa+gre,family=binomial)**" to answer this homework will be scored as 0, although you may use it to check your own answers.**

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -6 + gpa_i + 0.005gre_i$$

$$\frac{\pi_i}{1-\pi_i} = \exp\left(-6 + gpa_i + 0.005gre_i\right)$$

$$\pi_i = \left(1-\pi_i\right)\exp\left(-6 + gpa_i + 0.005gre_i\right) = \exp\left(-6 + gpa_i + 0.005gre_i\right) - \pi_i \exp\left(-6 + gpa_i + 0.005gre_i\right)$$

$$\pi_i\left(1 + \exp\left(-6 + gpa_i + 0.005gre_i\right)\right) = \exp\left(-6 + gpa_i + 0.005gre_i\right)$$

$$\pi_i = \frac{\exp\left(-6 + gpa_i + 0.005gre_i\right)}{\left(1 + \exp\left(-6 + gpa_i + 0.005gre_i\right)\right)}$$